Rozprawa doktorska

Wykrywanie zafałszowań, potwierdzanie autentyczności oraz identyfikacja zagrożeń biologicznych z wykorzystaniem chromatografii i modelowania chemometrycznego

mgr Barbara Krakowska

Promotor pracy:
dr hab. Michał Daszykowski, prof. UŚ

Instytut Chemii
Wydział Matematyki, Fizyki i Chemii
Uniwersytet Śląski

Składam serdeczne podziękowania

Panu dr hab. MICHAŁOWI DASZYKOWSKIEMU, prof. UŚ
za wszelką otrzymaną pomoc, wyrozumiałość i cierpliwość oraz cenne uwagi merytoryczne, które otrzymałam. Dziękuję również za możliwość przeprowadzenia badań i stworzenie wspaniałej atmosfery naukowej

Pani dr IVANIE STANIMIROVEJ-DASZYKOWSKIEJ oraz dr JOANNIE ORZEŁ za pomoc w prowadzonych badaniach, życzliwość i poświęcony czas

Dziękuję również za owocną współpracę

Panu dr IRENEUSZOWI GRABOWSKIEMU, Panu mgr MIROSŁAWOWI SZNAJDEROWI oraz Panu mgr GRZEGORZOWI ZALESZCZYKOWI z Izby Celnej w Białej Podlaskiej

Panu dr n. med. KAROLOWI FABIAŃCZYKOWI prezesowi firmy Polcargo International

Szczególnie dziękuję

MĘŻOWI i RODZICOM, którzy zawsze mnie wspierali i wierzyli we mnie

PRZYJACIOŁOM i ZNAJOMYM za życzliwość, cierpliwość i wsparcie w dążeniu do celu

Spis treści

Wykaz skrótów stosowanych w pracy

Skrót	Nazwa polska	Nazwa angielska
API	składnik aktywny leku	active pharmaceutical ingredient
AUC	pole powierzchni pod krzywą	area under curve
CART	drzewa klasyfikacji i regresji	classification and regression trees
CCR	procent poprawnej klasyfikacji	correct classification rate
COW	metoda zoptymalizowanego nakładania sygnałów maksymalizująca ich wzajemną korelację	correlation optimized warping
DAD	detektor z matrycą diodową	diode array detector
ED-XRF	spektroskopia rentgenowska z dyspersją energii	energy-dispersive X-ray spectroscopy
ELSD	detektor rozpraszania światła przez odparowanie	evaporative light scattering detector
FAME	estry metylowe kwasów tłuszczowych	fatty acid methyl esters
FN	próbka fałszywie negatywna	false negative
FP	próbka fałszywie pozytywna	false positive
GC-FID	chromatografia gazowa z detekcją płomieniowo-jonizacyjną	gas chromatography with flame ionization detector
GMP	dobra praktyka produkcyjna	good manufacturing practice
HPLC-DAD	wysokosprawna chromatografia cieczowa z detektorem z matrycą diodową	high-performance liquid chromatography with diode-array detector
kNN	metoda k-najbliższych sąsiadów	k-nearest neighbours

LDA	liniowa analiza dyskryminacyjna	linear discriminant analysis
MS	spektrometria mas	mass spectrometry
NCD	detektor chemiluminescencji azotu	nitrogen chemiluminescence detector
NIR	spektroskopia bliskiej podczerwieni	near infrared
NMR	magnetyczny rezonans jądrowy	nuclear magnetic resonance
PAsLS	metoda asymetrycznych najmniejszych kwadratów z funkcją kary	penalized asymmetric least squares
PCA	analiza czynników głównych	principal component analysis
PLS-DA	dyskryminacyjny wariant metody częściowych najmniejszych kwadratów	partial least squares discriminant analysis
ROC	charakterystyka operacyjna odbiornika	receiver operating characteristic
SCD	detektor chemiluminescencji siarki	sulfur chemiluminescence detector
SE	czułość	sensitivity
SIMCA	metoda modelowania indywidualnych grup	soft independent modelling of class analogies
SMC	metoda korelacji wieloczynnikowej	significance multivariate correlation
SNV	transformacja SNV	standard normal variate
SP	specyficzność	specificity
SR	współczynnik selektywności	selectivity ratio
TBT	tributylocyna	tributyltin
TN	próbka prawdziwie negatywna	true negative

TP	próbka prawdziwie pozytywna	true positive
UVE	metoda eliminacji zmiennych nieistotnych	uninformative variable elimination
VIP	metoda zmiennych znaczących dla projekcji	variable importance in projection

Streszczenie

Autentyczność produktów w wielu przypadkach zależy od ich składu chemicznego. Dlatego też do analizy produktów pod kątem ich autentyczności wykorzystuje się sygnały instrumentalne, które zawierają duży zasób informacji na temat substancji zawartych w próbce i mogą być postrzegane jako chemiczne odciski palca. Tego typu sygnał jest definiowany jako charakterystyczny profil opisujący skład chemiczny analizowanej próbki najlepiej jak to możliwe. Wśród wielu technik instrumentalnych, podejścia chromatograficzne są bardzo dobrym narzędziem do rejestracji chemicznych odcisków palca ze względu na możliwość rozdziału składników mieszanin.

W ramach badań do analizy chromatograficznych odcisków palca opracowano z powodzeniem różnego rodzaju podejścia chemometryczne w celu weryfikacji autentyczności wybranych produktów (olej napędowy, Viagra®) oraz badania obecności tributylocyny w próbkach środowiskowych. Przed przystąpieniem do analizy chemometrycznej zastosowano metody wstępnego przygotowania danych uzyskując poprawę jakości analizowanych sygnałów instrumentalnych. Następnie, zaproponowano modele diagnostyczne pozwalające przyporządkować badane próbki do rozważanych grup na podstawie chromatograficznych odcisków palca wykorzystując dyskryminacyjny wariant metody częściowych najmniejszych kwadratów, PLS-DA. Każdy model został poddany ocenie i opisany przez wybrane parametry walidacyjne charakteryzujące poprawność jego działania. Dodatkowo, w ramach prowadzonych badań zaproponowano nową procedurę konstrukcji i walidacji modeli diagnostycznych, która pozwala na jednoczesną estymację parametrów walidacyjnych modeli o różnej liczbie czynników dla zbioru modelowego i zbiorów testowych (wewnętrznego i zewnętrznego). Podejście to umożliwia uwzględnienie różnego rodzaju metod wyboru zmiennych istotnych na etapie budowy modelu PLS-DA, a tym samym wyznaczyć te zmienne (obszary chromatogramu), które są istotne dla rozróżniania analizowanych próbek. Ze względu na dobrą efektywność modeli diagnostycznych, opracowanych w celu weryfikacji autentyczności wybranych produktów i oceny zagrożenia biologicznego wynikającego z obecności substancji szkodliwych w próbkach środowiskowych, można wnioskować, że proponowane rozwiązania problemów badawczych z uwzględnieniem metod chemometrycznych mogą być z powodzeniem implementowane na potrzeby rutynowych analiz.

1. Wstęp

Obecnie, rynek produktów fałszowanych rozwija się na szeroką skale. Jest to głównie spowodowane niższą ceną takich produktów. Fałszowanie produktów definiuje się jako celową ingerencję człowieka w ich skład, wygląd lub procedurę wytwarzania. W zależności od obiektu fałszowania obserwuje się różne jego skutki. Gdy mamy do czynienia z fałszowaniem paliwa polegającym na usunięciu z niego dodatków akcyzowych, na szkodę narażony jest przede wszystkim budżet Państwa poprzez zaniżenie wpływów z tytułu należnego podatku akcyzowego [1]. Natomiast, gdy problem fałszowania dotyczy leków, stawka jest dużo wyższa, gdyż zagrożone jest zdrowie i życie ludzkie. Nielegalne wytwarzanie leków najczęściej odbywa się w prymitywnych warunkach niespełniających podstawowych norm czystości, a wytwarzane produkty są pozbawione kontroli jakości. Największym zagrożeniem w takim przypadku nie jest zaniżona zawartość substancji czynnej leku (co jest często obserwowane), a zanieczyszczenia pochodzące z substancji użytych do produkcji [2].

Pod pojęciem zafałszowania rozumiemy także domieszkowanie produktów tańszymi substancjami o podobnych właściwościach. Przykładem może być dodawanie do miodu syropu kukurydzianego w celu zwiększenia jego objętości [3]. Takie działanie jest nielegalne i bezpośrednio działa na szkodę konsumenta.

Przytoczone powyżej przykłady świadczą o dużej potrzebie kontrolowania parametrów jakości produktów, gdyż mają one wymiar finansowy, a także mogą oddziaływać na zdrowie i życie ludzi. Autentyczność produktu jest najczęściej związana z jego składem chemicznym (jakościowym i/lub ilościowym), jak również może być utożsamiana z pochodzeniem geograficznym [4]. W każdym z przypadków określa ona zgodność określonych cech danego produktu z deklaracją producenta. Odrębnym obszarem kontroli jakości jest analiza zanieczyszczeń środowiskowych, które podobnie jak zanieczyszczenia leków mogą oddziaływać na zdrowie i życie ludzi. Obecność w ekosystemie substancji zagrażających zdrowiu człowieka wymaga nie tylko ich stałej kontroli, ale również ciągłego ulepszania stosowanych metod analitycznych, co pozwala na wykrywanie coraz niższych stężeń analizowanych substancji. Spowodowane jest to koniecznością przestrzegania określonych norm definiujących dopuszczalne zawartości substancji szkodliwych w próbkach. Badania próbek o złożonym składzie jakimi są m.in. próbki żywności czy próbki środowiskowe, to tylko jedno z wyzwań analizy

jakościowej i ilościowej. Sygnały instrumentalne, posiadające duży zasób informacji o składzie chemicznym próbek mogą być traktowane jako tzw. chemiczne odciski palca. Analiza tego typu danych polega na porównaniu sygnałów instrumentalnych pomiędzy sobą lub względem sygnałów próbek referencyjnych. Takie podejście sprawdza się w przypadku oceny autentyczności, ponieważ często jej wyznacznikiem jest całościowy skład chemiczny analizowanego produktu. Złożoność sygnałów analitycznych wynika z sumowania się informacji pochodzących od poszczególnych komponentów próbki. W celu uzyskania optymalnego lub sub-optymalnego rozdziału chromatograficznego, który daje możliwość uzyskania istotnej informacji o składzie analizowanej próbki należy uprzednio dobrać warunki analizy m.in. kolumnę chromatograficzną, skład fazy ruchomej, warunki rozdziału. Dla próbek pochodzenia naturalnego, ze względu na ich złożony skład, uzyskanie optymalnego rozdziału chromatograficznego bywa bardzo trudne, a niejednokrotnie nawet niemożliwe. Jednym ze sposobów poprawy jakości sygnału analitycznego jest zastosowanie odpowiedniej procedury laboratoryjnej poprzedzającej rozdział chromatograficzny jak np. wstępne oczyszczanie próbki, jej zatężenie czy ekstrakcja. Poprawa jakości sygnału analitycznego wynikająca z zastosowanej techniki chromatograficznej może następować na skutek zwiększenia rozdzielczości (zastosowanie odpowiednich kolumn i rozpuszczalników) oraz poprzez wykorzystanie zaawansowanych detektorów takich jak np. spektrometr mas. Wstępne przygotowanie danych z wykorzystaniem technik matematycznych jest także sposobem na poprawę jakości sygnału np. poprzez eliminację linii podstawowej, nakładanie sygnałów czy usuwanie szumu.

Zastosowanie nowoczesnych technik instrumentalnych prowadzi do uzyskania dużej ilości danych, które mogą być trudne w interpretacji. W tym celu wykorzystywane są metody chemometryczne, które pozwalają na ekstrakcję użytecznej informacji ułatwiając tym samym, interpretację uzyskanych wyników analizy. W związku z tym różne podejścia chemometryczne znajdują coraz szersze zastosowanie do analizy całych sygnałów instrumentalnych stanowiących chemiczne odciski palca próbek w kontekście kontroli autentyczności wybranych produktów i oceny zagrożenia środowiskowego [5].

Niniejsza rozprawa doktorska obejmuje cykl badań, które zostały przedstawione w czterech publikacjach. Zaproponowałam w nich podejścia chemometryczne do oceny autentyczności wybranych produktów (olej napędowy, Viagra®) oraz weryfikacji obecności tributylocyny w wodzie na podstawie chromatograficznych odcisków palca.

Dodatkowo, opracowałam nowe podejście do konstrukcji i walidacji modeli dyskryminacyjnych bazujące na procedurze Monte Carlo. Publikacje wchodzące w skład rozprawy doktorskiej stanowią Załączniki nr 1-4 zamieszczone na końcu pracy.

[1] Detection of discoloration in diesel fuel based on gas chromatographic fingerprints, Analytical and Bioanalytical Chemistry, 407 (2015) 1159-1170; IF = 3,125, 35 pkt.*

[2] The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles, Analyst, 141 (2016) 1060-1070; IF = 4,033, 40 pkt.*

[3] Expert system for monitoring the tributyltin content in inland water samples, Chemometrics and Intelligent Laboratory Systems, 149 (2015) 123-131; IF = 2,217, 40 pkt.*

[4] Chemometrics and identification of counterfeit medicines – a review, Journal of Pharmaceutical and Biomedical Analysis, 127 (2016) 112-122; IF = 3,169, 35 pkt.*

* Punktacja zgodna z rokiem ukazania się publikacji według listy czasopism punktowanych MNiSW

2. Cele pracy

W ramach swojej pracy doktorskiej skupiłam się na następujących celach badawczych:

- ustalenie optymalnego zestawu metod chemometrycznych wykorzystywanych do ekstrakcji użytecznej informacji ze złożonych sygnałów chromatograficznych, w kontekście weryfikacji specyfikacji wybranych produktów;
- opracowanie wieloparametrowych modeli diagnostycznych wspomagających wykrywanie procederu odbarwiania paliw na podstawie chromatograficznych odcisków palca uzyskanych z wykorzystaniem chromatografii gazowej z detektorem płomieniowo-jonizacyjnym, (GC-FID);
- poszukiwanie obszarów sygnałów chromatograficznych, które różnicują grupy badanych próbek w kontekście badania autentyczności;
- potwierdzenie autentyczności preparatu Viagra® na podstawie chromatograficznych profili zanieczyszczeń;
- opracowanie i wykazanie użyteczności systemu eksperckiego bazującego na chromatograficznych odciskach palca poprzez ich modelowanie z wykorzystaniem wybranych metod uczenia maszynowego w celu oceny ryzyka skażenia wody tributylocyną i usprawnienia funkcjonowania laboratorium.

3. Część teoretyczna

3.1 Chromatograficzne odciski palca

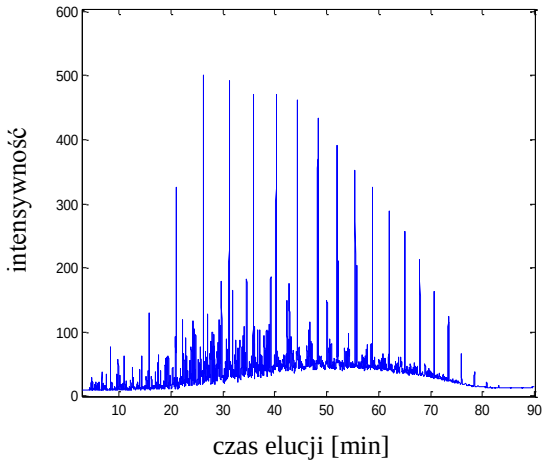
Analiza chromatograficzna polega na rozdziale składników próbki ze względu na ich powinowactwo do fazy stacjonarnej. Jest ona jedną z najczęściej stosowanych technik analitycznych wykorzystywanych do identyfikacji składników i analizy ilościowej złożonych próbek. Wynika to z możliwości jednoczesnego oznaczania wielu składników próbki w trakcie jednego rozdziału chromatograficznego.

Dzięki połączeniu standardowych technik chromatograficznych z różnymi detektorami wielokanałowymi (np. detektor z matrycą diodową, z ang. *diode array detector* – DAD, lub spektrometr masowy, z ang. *mass spectrometry detector* – MS) otrzymano sprzężone techniki chromatograficzne. Dzięki nim możliwe jest uzyskanie informacji o czystości pików chromatograficznych oraz polepszenie identyfikacji związków zawartych w badanej próbce. Z kolei efektywny rozdział chromatograficzny pozwala uzyskać pełniejszą informację o jej składzie chemicznym. Jednakże, analiza jakościowa poszczególnych składników próbki jest zazwyczaj skomplikowana, kosztowna i czasochłonna. Do badania próbek o złożonym składzie chemicznym wymagany jest rozdział komponentów i identyfikacja poszczególnych substancji chemicznych, co jest skomplikowane, a czasami nawet niemożliwe. Z tego powodu, do analizy porównawczej próbek stosowane są często całe sygnały instrumentalne stanowiące chemiczne odciski palca badanych próbek (z ang. *chemical fingerprints*). Przykładowy chromatogram stanowiący chromatograficzny odciska palca uzyskany dla próbki oleju napędowego przedstawiono na Rys. 1.

Chemiczny odcisk palca definiowany jest jako charakterystyczny profil reprezentujący skład chemiczny próbki najlepiej jak to możliwe. Optymalny chromatograficzny odcisk palca to chromatogram o relatywnie dużej rozdzielczości pików. Ta definicja implikuje konieczność odpowiedniego doboru warunków rozdziału, które są konsekwentnie stosowane dla całego zbioru próbek [6]. Chemiczny odcisk palca może stanowić sygnał instrumentalny uzyskany bezpośrednio dla próbki lub dla jej ekstraktu. Wykorzystanie ekstrakcji jest najczęściej uwarunkowane obecnością zanieczyszczeń zawartych w próbce, które mogą utrudniać analizę lub koniecznością zateżania badanego analitu. Ważnym jest, aby analizę porównawczą próbek opisanych przez chemiczne odciski palca prowadzić dla sygnałów uzyskiwanych tą samą metodą i przy zachowaniu tych

samych warunków rozdziału. Często do rejestracji chemicznych odcisków palca są stosowane detektory selektywne takie jak np. detektor chemiluminescencji azotu (z ang. *nitrogen chemiluminescence detector*, NCD) lub detektor chemiluminescencji siarki (z ang. *sulfur chemiluminescence detector*, SCD). Pozwalają one na otrzymanie selektywnej informacji dotyczącej jedynie związków zawierających odpowiednio azot lub siarkę. Stanowi to duże ułatwienie w analizie złożonych próbek pod kątem związków zawierających te atomy w swojej budowie. Detektor NCD w połączeniu z chromatografią gazową GC-NCD jest wykorzystywany m.in. do analizy dodatków akcyzowych i ich przemian w oleju napędowym. Ze względu na złożoność tego typu próbek i relatywnie małe stężenie dodatków akcyzowych w porównaniu do pozostałych składników paliwa, detektor NCD pozwala uzyskać selektywne sygnały instrumentalne zawierające informacje tylko o wybranej grupie związków.

Techniki chromatograficzne to doskonałe narzędzia rejestracji chemicznych odcisków palca. W odróżnieniu od technik spektroskopowych, interpretacja chromatograficznych odcisków palca jest łatwiejsza, gdyż w przypadku dobrego rozdziału jeden pik obserwowany na chromatogramie odpowiada jednej porcji eluatu. W idealnej sytuacji porcja eluatu zawiera czysty składnik, co można potwierdzić wykorzystując detektory wielokanałowe takie jak np. DAD czy MS. Cecha ta pozwala na uznanie sygnałów chromatograficznych jako unikalne źródło informacji o składzie badanej próbki.



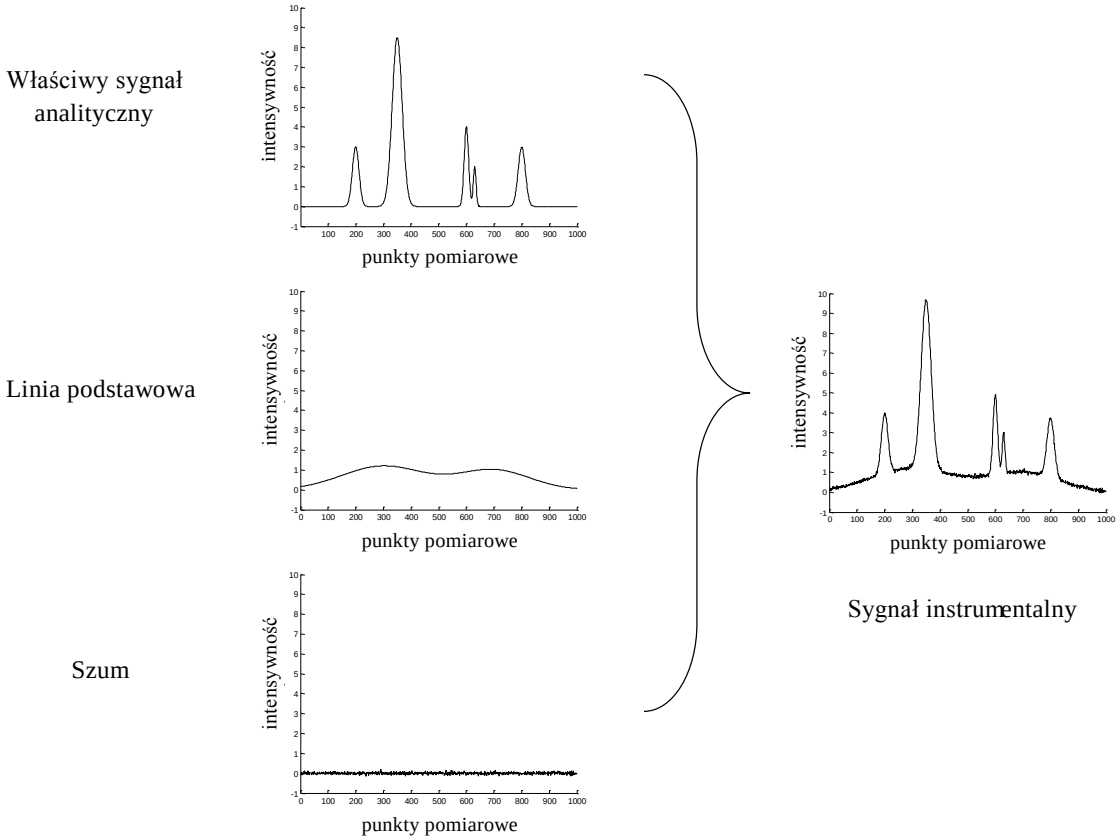
Rys. 1 Przykładowy chromatogram stanowiący chromatograficzny odciska palca próbki oleju napędowego, zarejestrowany za pomocą chromatografii gazowej z detektorem płomieniowo-jonizacyjnym

3.2 Chemometryczna analiza chromatograficznych odcisków palca

Chromatograficzne odciski palca, modelowane z wykorzystaniem narzędzi chemometrycznych jakimi są np. techniki dyskryminacyjne, pozwalają na konstruowanie modeli diagnostycznych. Znajdują one zastosowanie m.in. w badaniach autentyczności wybranych produktów czy do oceny zawartości substancji niebezpiecznych w próbkach środowiskowych. Dodatkowo, metody chemometryczne pozwalają na analizę sygnałów, dla których nie uzyskano optymalnego rozdziału chromatograficznego tzn., gdy niektóre piki nie są całkowicie od siebie oddzielone. Analiza chemometryczna chromatograficznych odcisków palca może być trudna ze względu na obecność dodatkowych komponentów sygnału takich jak szum, linia podstawowa czy przesunięcia pików. W zależności od ich udziału, uzyskane wyniki analizy surowych sygnałów instrumentalnych (chromatograficznych odcisków palca) mogą być niewiarygodne. Dlatego, przed przystąpieniem do modelowania danych chromatograficznych bardzo ważnym etapem jest ich wstępne przygotowanie. Polega ono na eliminacji czynników wpływających na jakość informacji takich jak szum, linia podstawowa czy przesunięcia odpowiadających sobie pików.

3.2.1 Wstępne przygotowanie chromatograficznych odcisków palca

Wszystkie sygnały instrumentalne, w tym także chromatograficzne odciski palca, składają się z trzech komponentów: szumu, linii podstawowej oraz pików pochodzących od komponentów próbki (zob. Rys. 2) [7]. Piki chromatograficzne opisują skład próbki zarówno pod względem ilościowym jak i jakościowym. Natomiast linia podstawowa i szum to komponenty sygnału będące skutkiem błędów pomiarowych i/lub niestabilności warunków prowadzenia rozdziału. Szum i linia podstawowa wnoszą dodatkową, niepożądaną zmienność do sygnału analitycznego zniekształcając piki pochodzące od poszczególnych składników. Tego typu zakłócenia sygnałów analitycznych mogą prowadzić do zafałszowania realnego obrazu składu próbki, a także utrudniają ich analizę porównawczą. Dzieje się tak, gdy linia podstawowa lub szum mają relatywnie dużą intensywność przez co przeprowadzenie zarówno analizy jakościowej jak i ilościowej jest trudne, a czasem wręcz niemożliwe.



Rys. 2 Elementy składowe sygnału instrumentalnego (właściwy sygnał analityczny, linia podstawowa, szum) na przykładzie chromatogramu opisującego pięcioskładnikową mieszaninę

Do najczęściej stosowanych metod wstępnego przygotowania sygnałów instrumentalnych zalicza się metody normalizacji, metody poprawiające stosunek sygnału do szumu oraz metody eliminujące przesunięcia pików względem siebie [7].

W pierwszym kroku przygotowania danych do analizy należy ocenić jakość sygnałów np. poprzez określenie stosunku sygnału do szumu lub wizualne zweryfikowanie intensywności takich komponentów jak szum czy linia podstawowa. Jedną z transformacji wykorzystywanych do wstępnego przygotowania danych jest normalizacja sygnałów. Stosuje się ją w celu umożliwienia porównania ze sobą sygnałów i polega na eliminacji błędów systematycznych, które występują w sygnałach m.in. z powodu niestabilności parametrów pobierania i przygotowywania próbek jak również nawet niewielkich wahań warunków prowadzenia analizy (np. różne objętości próbki nastrzykiwane na kolumnę chromatograficzną). Normalizacja polega na podzieleniu każdego elementu sygnału instrumentalnego przez określony parametr,

którego dobór zależy od stosowanego sposobu normalizacji. Najczęściej stosowana jest normalizacja sygnału do długości jednostkowej, polegająca na podzieleniu każdego elementu wektora przez pierwiastek sumy kwadratów jego wszystkich elementów. Inne warianty to normalizacja do jednostkowego pola powierzchni pod sygnałem oraz normalizacja SNV (z ang. *standard normal variate*) [8].

Szum jest definiowany jako odchylenie standardowe od wartości średniej sygnału rejestrowanego przez dany przyrząd pomiarowy. Charakteryzuje go wielkość określająca stosunek sygnału do szumu (stosunek średniej z sygnału do jego odchylenia standardowego) [9]. Szum jest komponentem sygnału o największej częstotliwości. Jego obecność jest uwarunkowana ograniczoną czułością stosowanego detektora oraz możliwością występowania w trakcie analizy reakcji pomiędzy składnikami zawartymi w próbce. Tego typu zjawiska mogą powodować zmiany natężenia sygnału instrumentalnego. Wyróżnia się kilka rodzajów szumu, między innymi tzw. szum biały o rozkładzie gaussowskim, szum skorelowany oraz szum proporcjonalny do sygnału. Szum ze względu na to, że nie wnosi istotnej informacji analitycznej, może negatywnie wpływać na dalszą analizę danych instrumentalnych. Można go wyeliminować poprzez zastosowanie różnego rodzaju filtrów takich jak na przykład filtr bazujący na medianie, filtr wykorzystujący wartość średnią sygnału, filtr Whitakkera lub filtr Savitzkyego-Golaya [9–11]. W zależności od sygnału analitycznego szum może być również korygowany za pomocą transformacji falkowej [12].

Kolejnym składnikiem chromatograficznych odcisków palca nie zawierającym informacji o składzie próbki jest linia podstawowa. Ma ona najmniejszą częstotliwość spośród składników sygnału. Z analitycznego punktu widzenia jest to sygnał instrumentalny zarejestrowany dla próbki pozbawionej badanych analitów. Kształt linii podstawowej jest zmienny i nawet dla zestawu próbek o tym samym pochodzeniu może się znacznie różnić. Dlatego usunięcie linii podstawowej jest istotnym krokiem wykonywanym przed analizą chemometryczną. Podobnie jak w przypadku szumu, intensywna linia podstawowa może powodować zafałszowanie wyników uzyskanych za pomocą wybranych metod chemometrycznych. W celu eliminacji linii podstawowej opracowano wiele metod, jednak do najczęściej stosowanych należy metoda asymetrycznych najmniejszych kwadratów z funkcją kary (z ang. *penalized asymmetric least squares*, PAsLS) [13]. Metoda PAsLS była stosowana do korekcji sygnałów chromatograficznych uzyskanych podczas badań realizowanych w ramach niniejszej

rozprawy doktorskiej. Wpływ linii podstawowej może być także eliminowany poprzez zastosowanie pochodnych sygnału.

Innym zjawiskiem, które utrudnia analizę porównawczą zbioru sygnałów instrumentalnych są przesunięcia pomiędzy pikami pochodzącymi od tych samych substancji. Główną przyczyną ich występowania jest zazwyczaj niestabilność warunków w trakcie prowadzonego rozdziału. W przypadku technik chromatograficznych ta niestabilność dotyczy m.in. starzenia się złoża kolumny chromatograficznej oraz fluktuacji składu fazy ruchomej. Jest to szczególnie niekorzystne zjawisko w przypadku gdy wykorzystywane są metody bazujące na porównywaniu ze sobą chemicznych odcisków palca. W takiej sytuacji pik pochodzący od tej samej substancji różni się położeniem na poszczególnych chromatogramach i wyniki przeprowadzonej analizy będą błędne, gdyż pomimo podobieństwa składu próbek mogą one zostać zidentyfikowane jako różniące się ze względu na zawartość/obecność danego komponentu. W celu usunięcia przesunięć pomiędzy pikami stosuje się techniki znane jako metody nakładania pików np. metodę zoptymalizowanego nakładania widm, która maksymalizuje wzajemną korelację sygnałów (z ang. *correlation optimized warping*, COW) [14,15].

Przed przystąpieniem do wstępnego przygotowania sygnałów instrumentalnych należy (jeśli jest to konieczne) zapewnić, tę samą liczbę punktów pomiarowych i częstotliwość próbkowania. Jest to wymagane w przypadku, gdy sygnały rejestrowane dla analizowanych próbek mają różną liczbę punktów pomiarowych, gdyż wówczas zestawienie ich w macierz jest niemożliwe.

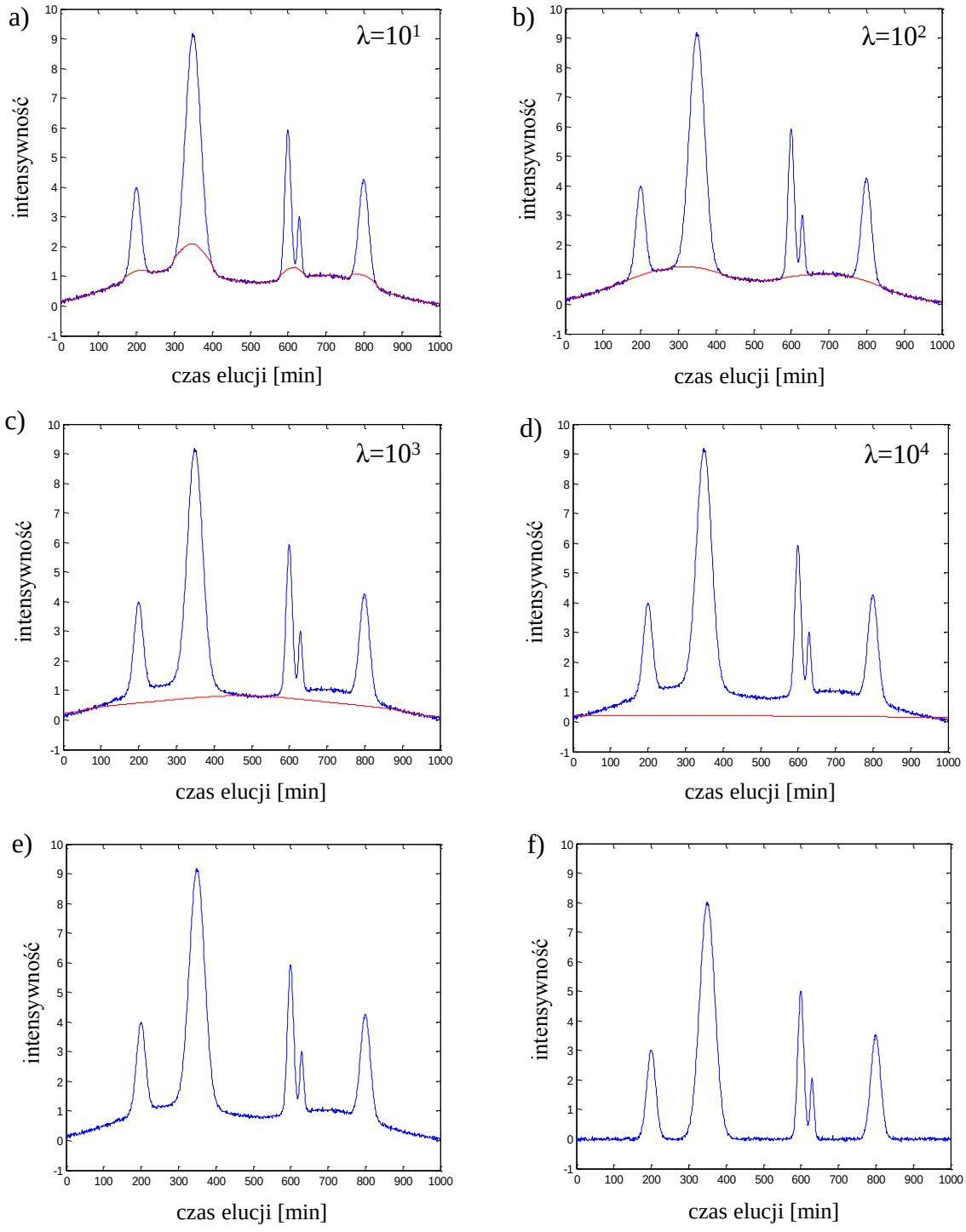
Eliminacja linii podstawowej

Jedną z najpopularniejszych metod stosowanych do eliminacji linii podstawowej jest metoda asymetrycznych najmniejszych kwadratów z funkcją kary, PAsLS (z ang. penalized asymmetric least squares) [13]. Poprzez minimalizację funkcji Q opisanej równaniem (1) wyznaczana jest linia podstawowa $\hat{y}$.

$$Q = \sum_i p_i(y_i - \hat{y}_i)^2 + \lambda \sum_i(\Delta^2 \hat{y}_i)^2 \quad (1)$$

gdzie y_i jest i-tym punktem pomiarowym sygnału instrumentalnego (np. chromatogramu), $\hat{y}_i$ jest i-tym punktem pomiarowym estymowanej linii podstawowej, p_i opisuje wagi dla różnicy zdefiniowanej w pierwszym członie równania, natomiast λ opisuje parametr kary dla drugiego członu równania. Operator pochodnej zastosowany do estymacji linii bazowej, jest oznaczony jako Δ. Pierwszy człon równania opisuje kwadraty reszt uzyskanych po odjęciu od sygnału aproksymowanej linii podstawowej z uwzględnieniem wag jakie odpowiednio wnoszą. Tym samym odzwierciedla on dopasowanie trendu linii podstawowej do korygowanego sygnału instrumentalnego. Natomiast parametr λ w drugim członie równania dotyczy 'niewygładzonych' obszarów sygnału analitycznego, których kształt wykracza poza kształt estymowanej linii podstawowej (np. piki na chromatogramie).

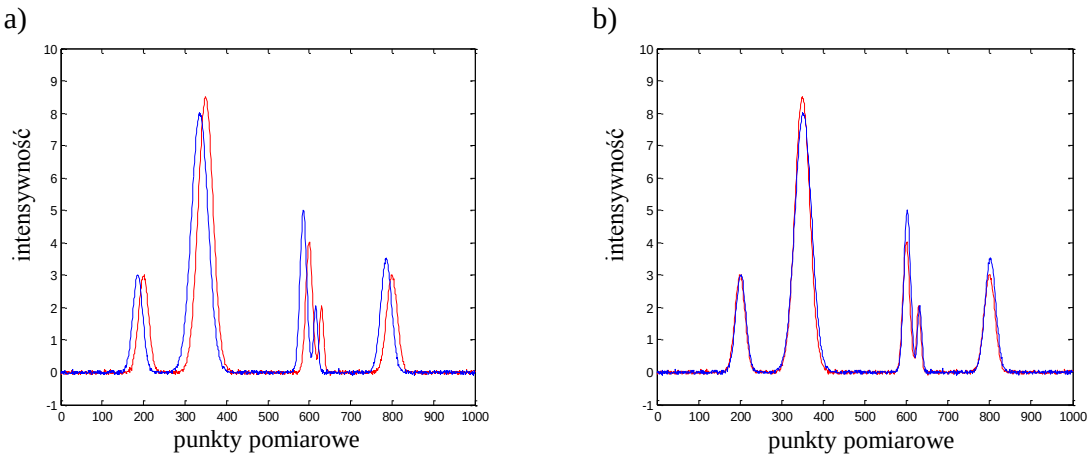
Estymacja linii podstawowej za pomocą metody PAsLS wymaga wyznaczenia dwóch parametrów wejściowych (λ, p), których wartości są dobierane poprzez wizualną ocenę działania metody dla wielu zestawów parametrów dokonywaną przez użytkownika. Parametr p definiuje asymetryczność danego sygnału instrumentalnego, natomiast parametr λ jest związany z stopniem wygładzenia linii podstawowej. W literaturze podawane są przedziały wartości danych parametrów, w których należałoby szukać optymalnego zestawu ich wartości i wynoszą odpowiednio $10^{-3} \leq p \leq 10^{-1}$ oraz $10^2 \leq \lambda \leq 10^9$. Poniżej przedstawiono przykładowe wyniki estymacji linii podstawowej dla różnych wartości parametru λ oraz przykładowy sygnał instrumentalny przed i po skorygowaniu linii podstawowej.



Rys. 3 (a-d) estymacja linii podstawowej dla różnych wartości parametru λ, (e) oryginalny sygnał oraz (f) sygnał po usunięciu linii podstawowej za pomocą metody asymetrycznych najmniejszych kwadratów z funkcją kary dla λ = 100

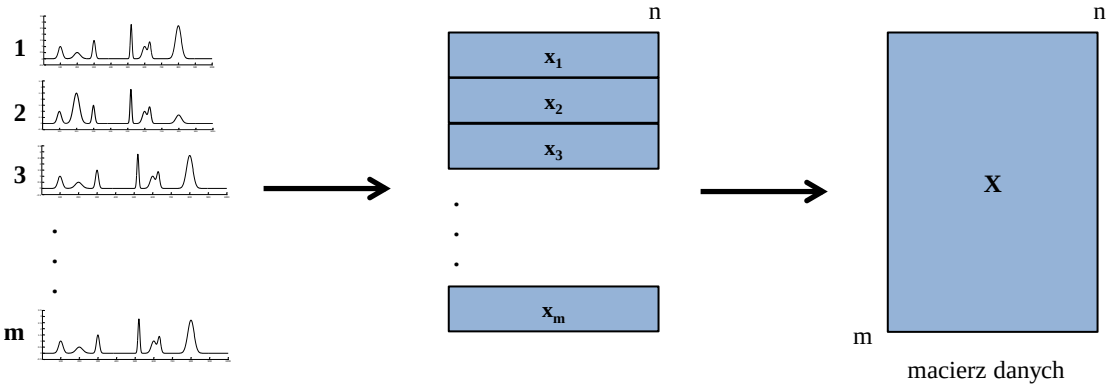
Eliminacja przesunięć pików chromatograficznych

Obecność przesunięć pików w sygnałach chromatograficznych jest częstym zjawiskiem spowodowanym niestabilnością warunków analizy. Aby było możliwe porównanie sygnałów ze sobą należy nałożyć na siebie odpowiadające sobie piki tj. piki pochodzące od tych samych substancji. W tym celu stosuje się wiele metod m.in. metodę zoptymalizowanego nakładania sygnałów maksymalizującą ich wzajemną korelację (z ang. *correlaction optimizing warping*, COW) [14], metodę automatycznego nakładania sygnałów (z ang. *automatic alignment*, AA) [16] oraz metodę nakładania sygnałów z zastosowaniem logiki rozmytej (z ang. *fuzzy warping*) [17]. Najwięcej aplikacji ma metoda COW, która dokonuje korekcji przesunięć pików w sygnałach względem sygnału wzorcowego. Istnieje kilka sposobów wyboru sygnału wzorcowego [15]. Jednym z nich jest podejście bazujące na współczynniku korelacji pomiędzy analizowanymi sygnałami, zgodnie z którym dla każdego sygnału ze zbioru danych wyznaczane są współczynniki korelacji z pozostałymi sygnałami. Następnie uzyskane wartości współczynników korelacji dla pojedynczego sygnału są uśredniane, a jako sygnał wzorcowy wybierany jest ten, który charakteryzuje się największą średnią wartością współczynników korelacji. Dopasowanie sygnałów do sygnału wzorcowego jest uzyskiwane poprzez liniową interpolację prowadzącą do rozciągania i/lub kompresji poszczególnych fragmentów sygnału w taki sposób, aby uzyskać ich maksymalną korelację z sygnałem wzorcowym. W metodzie COW jakość nakładania sygnałów zależy od dwóch parametrów. Pierwszy, oznaczany symbolem N, charakteryzuje liczbę segmentów na którą będzie podzielony każdy sygnał. Drugi to tzw. parametr elastyczności s, który definiuje możliwe położenia końców poszczególnych sekcji, na które został podzielony sygnał. Im większa wartość parametru elastyczności tym zdolność kompensowania przesunięć pików wzrasta. Z reguły, podczas procesu nakładania testuje się szereg kombinacji parametrów N i s, a jako optymalne wybierane są te, które pozwalają uzyskać maksymalną wartość korelacji sygnału z sygnałem wzorcowym. Wyniki działania metody COW dla dwóch przykładowych chromatogramów oraz N = 20, s = 3 przedstawiono na Rys. 4. Wartości współczynnika korelacji pomiędzy sygnałami przed i po ich nałożeniu wynoszą odpowiednio 0,724 i 0,992.



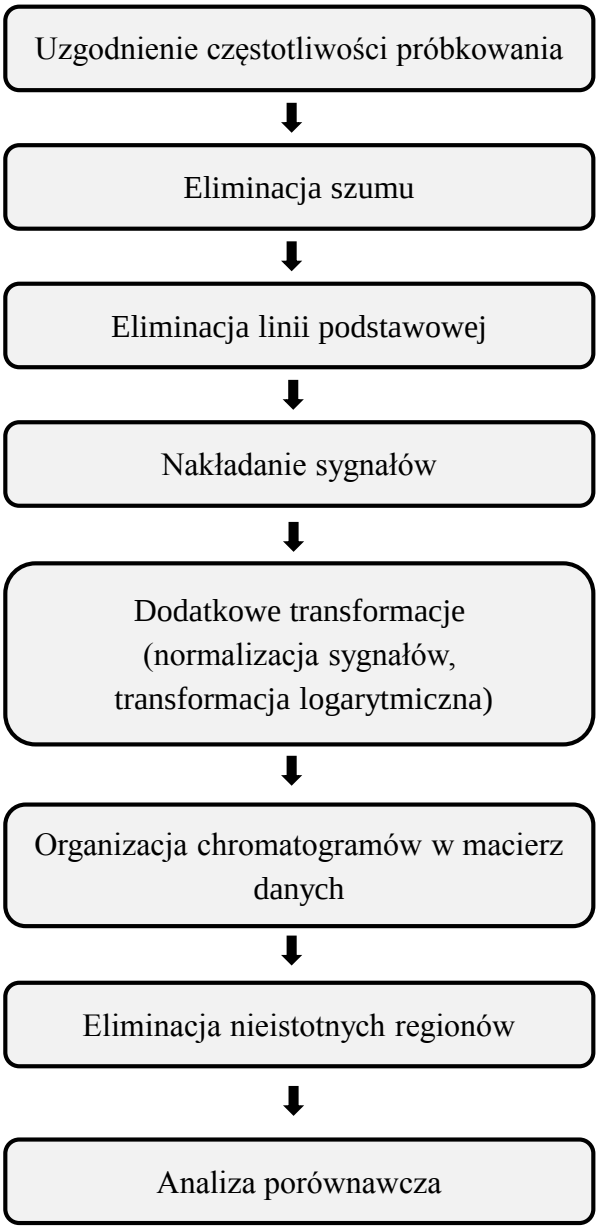
Rys. 4 Nakładanie dwóch przykładowych sygnałów chromatograficznych metodą zoptymalizowanego nakładania sygnałów maksymalizującą ich wzajemną korelację, (a) oryginalne sygnały przed nałożeniem pików, (b) sygnały po eliminacji przesunięć pomiędzy pikami (niebieska linia - sygnał wzorcowy, czerwona linia - sygnał korygowany)

Wstępnie przygotowane sygnały instrumentalne organizowane są w macierz danych, **X**, w przypadku gdy sygnały mają postać wektora lub w tensor, gdy pojedynczą próbkę opisuje zbiór sygnałów stanowiący macierz danych np. HPLC-DAD. W przypadku sygnałów instrumentalnych, które dla każdej próbki mają postać wektora tak jak np. chromatogramy rejestrowane za pomocą detektorów jednokanałowych (np. GC-FID) czy widma NIR, organizuje się je w macierz danych w taki sposób, że każdy z sygnałów reprezentujący badaną próbkę stanowi kolejny wiersz macierzy. Schemat organizacji sygnałów chromatograficznych w macierz danych przedstawia Rys. 5.



Rys. 5 Schemat organizacji sygnałów instrumentalnych w macierz danych, **X**

Omówione zagadnienia dotyczące wstępnego przygotowania danych instrumentalnych są typowe dla wszystkich rodzajów sygnałów analitycznych. Ich umiejętne zastosowanie pozwala w dużej mierze eliminować niepożądane źródła wariancji, a przez to uzyskać lepsze rezultaty analizy porównawczej i modelowania danych. Ogólny schemat kolejnych kroków przygotowania sygnałów chromatograficznych do analizy chemometrycznej przedstawia Rys. 6.



Rys. 6 Główne kroki przygotowania chromatograficznych odcisków palca do analizy chemometrycznej

3.2.2 Metody chemometryczne stosowane do badania autentyczności próbek

Metody chemometryczne są coraz częściej wykorzystywane do analizy chemicznych odcisków palca w kontekście badania autentyczności próbek oraz oceny zagrożeń biologicznych spowodowanych obecnością substancji niebezpiecznych w próbkach środowiskowych. W zależności od podjętego problemu badawczego, stosuje się różne podejścia chemometryczne. Są to zarówno metody, które ułatwiają interpretację i ekstrakcję informacji zawartych w danych eksperymentalnych, jak również pozwalające na budowę reguł logicznych wspierających rozróżnienie analizowanych grup próbek i prognozowanie ich przynależności do odpowiednich grup. Ogólnie metody chemometryczne stosowane do weryfikacji jakości różnego rodzaju produktów można podzielić na trzy grupy: metody eksploracyjne, klasyfikacyjne oraz dyskryminacyjne.

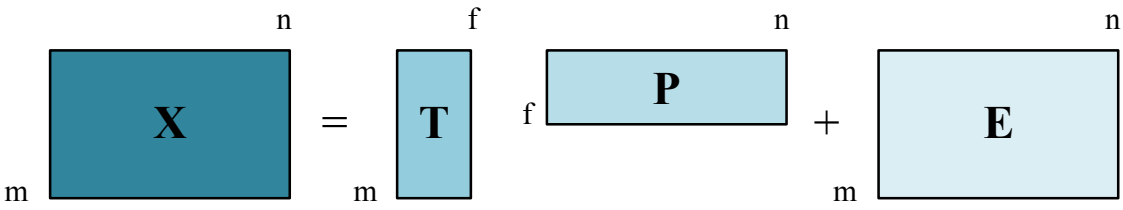
Metody eksploracyjne

Metody eksploracyjne należą do metod uczenia bez nadzoru. Mają na celu ujawienie struktury danych, a w szczególności grupowania się obiektów o podobnych właściwościach próbek znacznie różniących się od pozostałych czy lokalnych fluktuacji gęstości danych. Do typowych technik uczenia bez nadzoru należy analiza czynników głównych (z ang. *principal component analysis*, PCA) [18,19].

Celem metody PCA jest modelowanie, kompresja i wizualizacja wielowymiarowych danych. W analizie eksploracyjnej z wykorzystaniem tej metody macierz danych **X** o m obiektach i n parametrach jest przedstawiona jako iloczyn dwóch macierzy **T** i **P** o wymiarowości odpowiednio [m,f] i [f,n] (zob. równanie 2). Macierz **T** zawiera współrzędne obiektów (tzw. wyniki), a macierz **P** współrzędne parametrów dla nowych ukrytych zmiennych tzw. czynników głównych. Graficzna postać dekompozycji danych z zastosowaniem modelu PCA została przedstawiona na Rys. 7.

$$\mathbf{X}_{[m,n]} = \mathbf{T}_{[m,f]}\,\mathbf{P}^{T}_{[f,n]} + \mathbf{E}_{[m,n]} \qquad (2)$$

gdzie, **X** to macierz wyjściowych danych, **T** to macierz wyników, **P** reprezentuje macierz wag, a **E** to macierz reszt od modelu, f oznacza liczbę czynników głównych, a m i n to odpowiednio liczba próbek i zmiennych.



Rys. 7 Dekompozycja macierzy danych **X** z wykorzystaniem modelu PCA z f czynnikami do macierzy wyników **T**, wag **P** oraz reszt **E**

Czynniki główne w metodzie PCA są konstruowane w sposób iteracyjny tak, aby maksymalizować opis wariancji danych. Każdy kolejny czynnik główny modeluje wariancję nieopisaną przez poprzednie czynniki główne. A zatem, wkład każdego kolejnego czynnika głównego do opisu całkowitej wariancji danych jest coraz mniejszy.

Wstępna ocena struktury danych za pomocą projekcji wyników i/lub wag pozwala określić zależności istniejące pomiędzy próbkami i/lub parametrami, jak również ułatwia interpretację wyników uzyskanych w kolejnych etapach analizy chemometrycznej. Główną zaletą analizy danych za pomocą metody PCA jest brak konieczności posiadania wiedzy na temat przynależności analizowanych danych do poszczególnych grup, co odróżnia metody uczenia bez nadzoru od metod uczenia z nadzorem.

Metody klasyfikacyjne

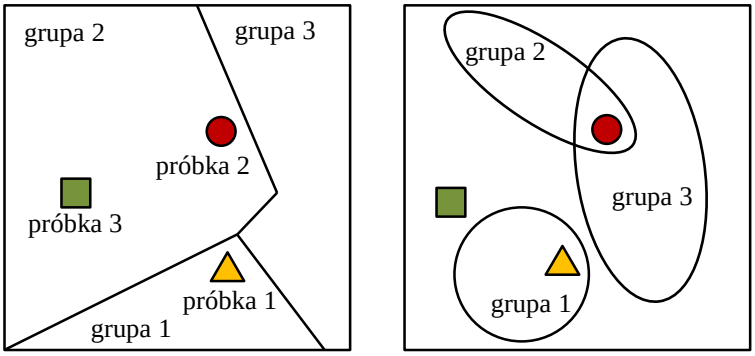
Metody klasyfikacyjne to metody uczenia z nadzorem, ponieważ wykorzystują do konstrukcji modelu zbiór danych eksperymentalnych **X** i zmienną zależną **y**. Zmienna zależna zawiera informację o przynależności próbki do danej grupy. Metody klasyfikacyjne zakładają, iż każda grupa próbek jest modelowana oddzielnie, a próbkę można przypisać do jednej z grup, do kilku grup jednocześnie, albo do żadnej z nich (Rys. 8b).

Najczęściej stosowaną metodą klasyfikacyjną jest metoda modelowania indywidualnych grup, SIMCA (z ang. *soft independent modelling of class analogies*) [18]. Metoda SIMCA buduje reguły klasyfikacyjne na podstawie parametrów modelu PCA otrzymanego oddzielnie dla każdej grupy próbek. Ustalenie przynależności próbki do danej grupy odbywa się poprzez ocenę jej odległości od próbek modelowych w przestrzeni modelu (odległość Mahalanobisa) oraz dopasowaniu próbki do modelu (reszty od modelu). W praktyce oznacza to, że przestrzeń modelu jest ograniczona przez obszar definiowany odpowiednio dobranymi wartościami progowymi. Określenie „soft" w metodzie SIMCA odnosi się do możliwości przypisania pojedynczej próbki do kilku grup jednocześnie.

Metody klasyfikacyjne są zazwyczaj wykorzystywane w sytuacjach gdy nie ma możliwości uwzględnienia wszystkich grup próbek na etapie budowy modelu. Wówczas minimalizowane jest ryzyko identyfikacji przez model próbek fałszywie pozytywnych, tj. takich, które są rozpoznawane jako należące do danej grupy podczas gdy nie powinny.

Metody dyskryminacyjne

Metody dyskryminacyjne to grupa metod uczenia z nadzorem, za pomocą których przestrzeń zmiennych eksperymentalnych zostaje podzielona na kilka wzajemnie wykluczających się podprzestrzeni. Ich liczba jest równa liczbie grup w rozpatrywanym problemie dyskryminacyjnym. Ze względu na położenie próbki w przestrzeni zmiennych objaśniających jest ona zawsze przypisana tylko do jednej grupy. Ta własność zasadniczo różni tę grupę metod od metod klasyfikacyjnych. Schematycznie różnice pomiędzy metodami dyskryminacyjnymi i klasyfikacyjnymi przedstawiono na Rys. 8.



przestrzeń zmiennych w metodach dyskryminacyjnych

przestrzeń zmiennych w metodach klasyfikacyjnych

Rys. 8 Ilustracja przewidywania przynależności do grup w technikach (a) dyskryminacyjnych i (b) klasyfikacyjnych

W przypadku problemu dwuklasowego przynależność próbek do analizowanych grup jest określana za pomocą zmiennej zależnej **y** mającej postać wektora. Dla modelu PLS-DA poszczególne elementy zmiennej **y** są definiowane za pomocą kodowania binarnego (0, 1) lub bipolarnego (-1, 1). Przyjęta etykieta dla danej grupy jest kwestią umowną. W przypadku kodowania binarnego wszystkie próbki, które na podstawie modelu dyskryminacyjnego otrzymają wartość zmiennej zależnej większą od 0,5 przypisane są do grupy oznaczonej za pomocą jedynek, natomiast wszystkie próbki dla których wartość zmiennej zależnej jest mniejsza od 0,5 są przypisane do grupy oznaczonej za pomocą zer. Do konstrukcji modeli dyskryminacyjnych stosowane są takie techniki chemometryczne jak, liniowa analiza dyskryminacyjna (z ang. *linear discriminant analysis*, LDA) [20], drzewa klasyfikacji i regresji (z ang. *classification and regression trees*, CART) [21], metoda k-najbliższych sąsiadów kNN (z ang. *k-nearest neighbors*) [22] oraz dyskryminacyjny wariant metody częściowych najmniejszych kwadratów (z ang. *discriminant partial least squares discriminant analysis*, PLS-DA) [23–25]. Metodę PLS-DA można przedstawić w postaci równania 3:

$$\mathbf{y}_{[m,1]} = \mathbf{X}_{[m,n]}\,\mathbf{b}^{T}_{[n,1]} + \mathbf{e}_{[m,1]} \qquad (3)$$

gdzie, **y** to wektor zmiennych zależnych, **X** to zbiór zmiennych objaśniających, **b** to wektor współczynników regresji maksymalizujących wariancję w macierzy

X i kowariancję pomiędzy macierzą **X,** a zmienną **y**, **e** jest wektorem reszt od modelu, a m i n to odpowiednio liczba próbek i zmiennych.

Zarówno badanie produktów pod względem ich autentyczności jak i ocena zgodności próbki z określoną normą stanowią problem dwuklasowy, ponieważ analizowana próbka może być albo autentyczna albo nie. Natomiast w przypadku oceny zagrożeń biologicznych spowodowanych obecnością substancji szkodliwych w próbkach środowiskowych zawartość badanych analitów w skażonej próbce przekracza lub nie przekracza dopuszczalnych stężeń co w pełni uzasadnia użycie metody dyskryminacyjnej PLS-DA w badaniach dotyczących niniejszej rozprawy doktorskiej.

Wybór zbioru modelowego i testowego

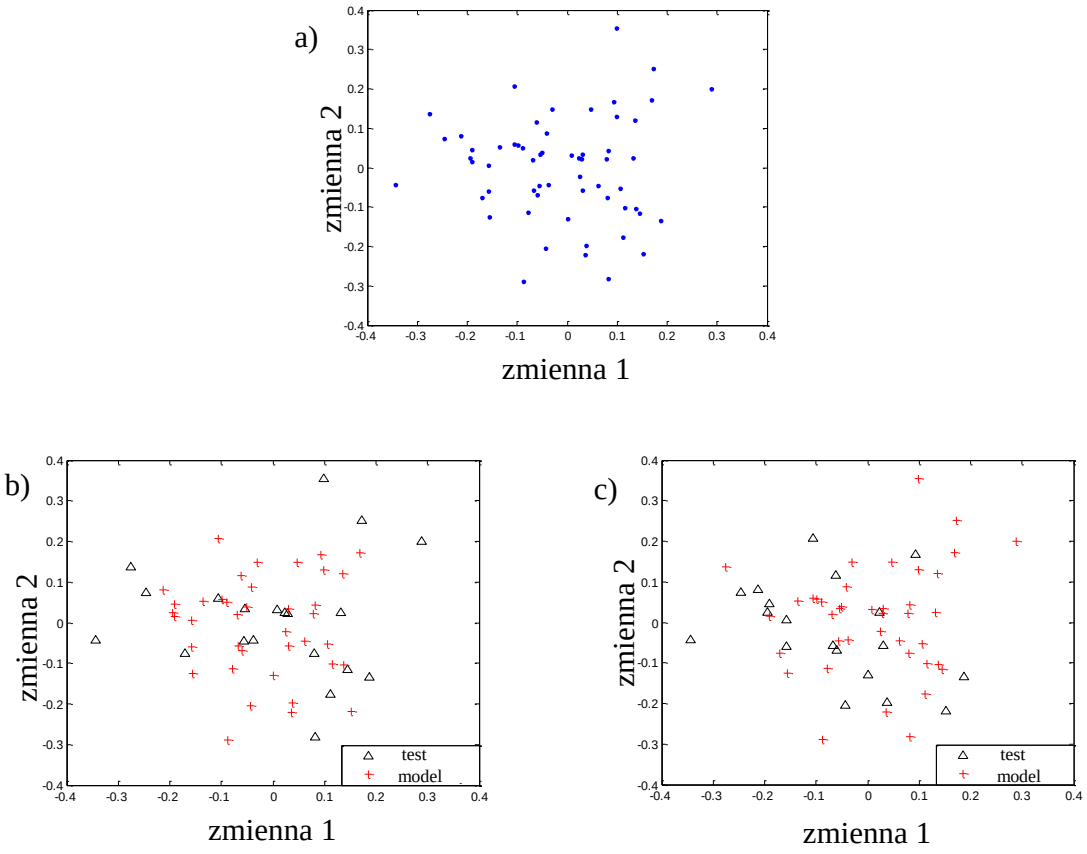
Konstrukcja modelu kalibracyjnego/dyskryminacyjnego wymaga użycia odpowiedniej liczby próbek modelowych, które są wykorzystywane do budowy reguł klasyfikacyjnych/dyskryminacyjnych. Aby zapewnić dobre zdolności predykcyjne modelu, zbiór modelowy powinien zawierać próbki reprezentujące wszystkie możliwe źródła wariancji danych, a więc takie, które pokrywają całą przestrzeń analizowanego zbioru danych. W przypadku gdy zbiór modelowy nie jest reprezentatywny, pojawia się ryzyko ekstrapolacji lub interpolacji modelu, co może skutkować pogorszeniem zdolności przewidywania. Reprezentatywność zbioru modelowego może być uzyskana m.in. poprzez odpowiednie zaplanowanie eksperymentu [26]. Jednak w niektórych sytuacjach, gdy obiektem badań są próbki środowiskowe lub próbki naturalne o nieznanym składzie (np. zafałszowane leki), użycie technik planowania eksperymentu jest niemożliwe. Taka sytuacja miała miejsce w przypadku przedstawionych badań. Wówczas w celu wyboru reprezentatywnych próbek z zestawu analizowanych danych stosowane są metody wyboru zbioru próbek takie jak metoda Duplex [27] oraz algorytm Kennarda i Stona [28]. Zapewniają one możliwie najlepszą reprezentatywność zbioru modelowego poprzez włączanie do niego próbek, które równomiernie pokrywają przestrzeń danych eksperymentalnych. Zarówno metoda Duplex jak i algorytm Kennarda i Stona mogą być stosowane, gdy liczba analizowanych próbek jest odpowiednio duża. Zazwyczaj przyjmuje się, że zbiór modelowy powinien zawierać od 70% do 75% całkowitej liczby próbek, natomiast pozostałe próbki tworzą zbiór testowy. W obu algorytmach podobieństwo pomiędzy próbkami jest wyrażone za pomocą

odległości euklidesowej. Dodatkowo bardzo ważne jest aby do konstrukcji modelu PLS-DA stosowany był zbilansowany zbiór modelowy co oznacza, że zbiór ten powinien być zbudowany z takiej samej liczby próbek z poszczególnych grup [29]. Niezbalansowany zbiór modelowy powoduje przesunięcie granicy dzielącej przestrzeń danych ze względu na przynależność do analizowanych grup w kierunku grupy bardziej licznej czego konsekwencją może być gorsze przewidywanie modelu.

W przypadku **algorytmu Kennarda i Stonea** pierwszą próbką wybraną do zbioru modelowego jest ta, która jest położona najbliżej arytmetycznego środka danych. Kolejną próbką wybraną do zbioru modelowego jest próbka znajdująca się najdalej od pierwszej. Jako trzecią do zbioru modelowego wybiera się próbkę najbardziej oddaloną od dwóch dotychczas wybranych. W analogiczny sposób wybiera się kolejne próbki do zbioru modelowego do momentu, gdy zbiór będzie zawierał ich założoną liczbę. Próbki, które nie zostały wybrane do zbioru modelowego stanowią zbiór testowy.

Algorytm „Duplex”, w odróżnieniu od algorytmu Kennarda i Stona, ma na celu zapewnienie reprezentatywności zarówno zbioru modelowego jak i testowego. W pierwszym kroku identyfikuje się dwie próbki najbardziej od siebie różne i włącza je do zbioru modelowego. Kolejna para próbek, która jest od siebie również najbardziej oddalona, jest dodana do zbioru testowego. W następnych krokach wybierane są naprzemiennie do zbioru modelowego i testowego kolejne pary próbek najbardziej od siebie oddalonych. Procedura wyboru próbek jest wykonywana aż do momentu, gdy do zbioru testowego zostanie przyporządkowana określona liczba próbek.

Przykładowy podział danych na zbiór modelowy i testowy za pomocą algorytmów Duplex i Kennarda i Stona przedstawiono na Rys. 9.



Rys. 9 (a) wizualizacja próbek na płaszczyźnie zdefiniowanej przez dwie zmienne, podział próbek na zbiór modelowy i testowy za pomocą algorytmów (b) Duplex i (c) Kennarda i Stona

Parametry walidacyjne

Znanych jest wiele parametrów walidacyjnych charakteryzujących efektywność działania modeli dyskryminacyjnych i klasyfikacyjnych. Są one obliczane niezależnie dla zbioru modelowego i testowego. Najbardziej popularnym parametrem oceny modelu jest procent poprawnej klasyfikacji (z ang. *correct classification rate*, CCR), który mówi o liczbie próbek, których przynależność do grup została właściwie rozpoznana przez model.

Innymi parametrami oceny modelu są np. czułość i specyficzność. Do ich obliczenia wykorzystuje się informację o liczbie próbek poprawnie lub niepoprawnie zaklasyfikowanych na podstawie modelu oddzielnie dla każdej z analizowanych grup lub w całym zbiorze próbek. Dla problemu dyskryminacyjnego, który dotyczy tylko dwóch grup próbek, tak jak ma to miejsce w problemach identyfikacji autentyczności,

czy oceny zgodności próbki z zakładaną normą zakłada się, że próbki autentyczne lub próbki spełniające określone normy stanowią grupę pozytywną, a zafałszowane czy niespełniające normy to grupa negatywna. Jako próbki prawdziwie pozytywne (z ang. *true positive*, TP) i prawdziwie negatywne (z ang. *true negative*, TN) uznaje się te, które są poprawnie przyporządkowane do danej grupy za pomocą modelu. Próbki fałszywie pozytywne (z ang. *false positive*, FP) są to próbki zafałszowane (negatywne) rozpoznawane przez model dyskryminacyjny jako autentyczne (pozytywne). Analogicznie, próbki fałszywie negatywne (z ang. *false negative*, FN) są to próbki autentyczne (pozytywne) błędnie przypisywane do grupy próbek zafałszowanych (negatywnych). Czułość (z ang. *sensitivity*, SE) dla danej grupy próbek definiuje się jako iloraz liczby próbek prawdziwie pozytywnych i liczby wszystkich próbek pozytywnych i mówi o tym jak dobrze dany model przewiduje próbki autentyczne. Poprawność przewidywania próbek negatywnych charakteryzowana jest przez specyficzność modelu (z ang. *specificity*, SP), określająca stosunek liczby próbek prawdziwie negatywnych do liczby wszystkich próbek negatywnych w analizowanym zbiorze danych. Wszystkie opisane parametry przedstawiają poniższe równania (4-6):

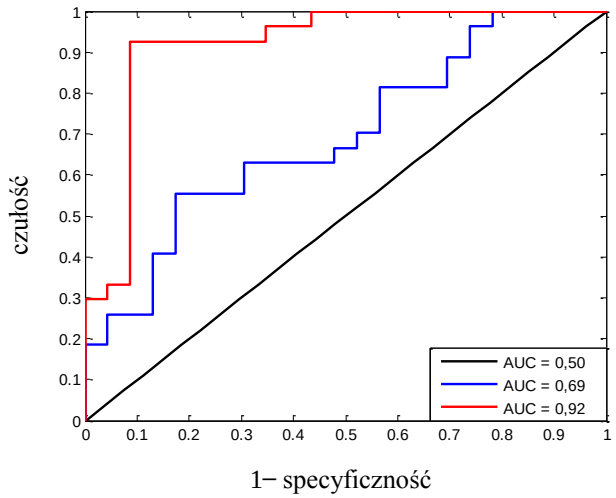
$$SE = TP / (TP + FN) \quad (4)$$

$$SP = TN / (TN + FP) \quad (5)$$

$$CCR = (TP + TN) / (TP + FP + FN + TN) \quad (6)$$

Kolejnym sposobem oceny jakości modelu dyskryminacyjnego jest analiza krzywej ROC (z ang. *receiver operating characteristic*) [30]. Obrazuje ona zależność pomiędzy procentem próbek prawdziwie pozytywnych i fałszywie pozytywnych. Im bardziej krzywa ma wypukły kształt tym model dyskryminacyjny jest bardziej wiarygodny. Krzywa ROC może być również opisana za pomocą pola powierzchni pod jej wykresem, tzw. parametr AUC (z ang. *area under curve*). Parametr ten obrazuje efektywność przewidywania modelu. Im bardziej wartość pola powierzchni pod krzywą ROC (AUC) jest zbliżona do 1 tym lepszą efektywność posiada dany model. Najlepszą dyskryminację próbek zapewnia model, którego wartość AUC wynosi 1. Gdy AUC wynosi 0,5 oznacza to, że dany model działa nie lepiej niż losowe przyporządkowywanie próbek do dwóch grup.

Przykładowe krzywe ROC dla modeli o różnych właściwościach predykcyjnych wraz z ich wartościami AUC przedstawia Rys. 10.



Rys. 10 Przykładowe krzywe ROC dla modeli skonstruowanych za pomocą dyskryminacyjnego wariantu metody częściowych najmniejszych kwadratów o różnych wartościach parametru AUC

Estymacja rozkładu wartości parametrów walidacyjnych

W celu uzyskania bardziej realistycznych estymacji rozkładu wartości parametrów opisujących konstruowany model można stosować różnego rodzaju podejścia. Do najczęściej wymienianych należą procedury ‘bootstrap’ jackknifing, kroswalidacja oraz Monte Carlo [31–35].

Bootstraping polega na wielokrotnym losowaniu ze zwracaniem określonej liczby próbek do zbioru modelowego. Metoda ta pozwala symulować jak wpływa zmienność zbioru modelowego na konstrukcję i poprawność przewidywania modelu diagnostycznego. Na podstawie każdego wylosowanego zbioru modelowego konstruowany jest model, a zbiór testowy tworzą próbki, które nie zostały włączone do zbioru modelowego i służą one do oceny właściwości predykcyjnych danego modelu.

Jackknifing polega na wyłączaniu ze zbioru danych pojedynczego obiektu, który następnie służy do testowania modelu skonstruowanego na podstawie pozostałych próbek. Daną procedurę prowadzi się wielokrotnie wyłączając z każdą iteracją kolejną próbkę ze zbioru danych oraz powtarzając konstrukcję oraz testowanie modelu. W tym

przypadku liczba możliwych iteracji jest ograniczona i odpowiada liczbie próbek znajdujących się w zbiorze danych.

Metoda kroswalidacji polega na podziale zbioru danych na wiele podzbiorów o określonej liczbie próbek (k). Każdy z podzbiorów jest wyłączany ze zbioru danych i stanowi zbiór testowy, natomiast pozostałe próbki służą do skonstruowania modelu diagnostycznego. Szczególnym typem kroswalidacji jest kroswalidacja „typu wyrzuć jeden obiekt" (k = 1), w tym przypadku w pojedynczej iteracji ze zbioru danych wyłączany jest jeden obiekt stanowiący jednocześnie zbiór testowy. Kroswalidacja typu „wyrzuć jeden obiekt" jest stosowana w przypadkach, gdy zbiór danych zawiera małą liczbę próbek.

Kolejnym sposobem estymacji zmienności zbioru modelowego jest metoda Monte Carlo. Polega ona na losowym podziale zbioru próbek na dwa podzbiory, który jest wykonywany wielokrotnie. Za każdym razem do zbioru testowego włączana jest ta sama liczba próbek stanowiąca od 30% do 50% całkowitej liczby próbek znajdujących się w zbiorze danych.

Dzięki wymienionym podejściom uzyskuje się rozkład wybranych parametrów walidacyjnych opisany przez ich wartość średnią i odchylenie standardowe, co pozwala wyznaczyć zakresy niepewności ich oszacowania.

Metody wyboru zmiennych

Często modele diagnostyczne są konstruowane na podstawie danych zawierających znacznie większą liczbę parametrów w stosunku do liczby próbek. Ta sytuacja zwiększa ryzyko przeuczenia modelu. Zjawisko to polega na dopasowywaniu modelu zarówno do danych jak i do przypadkowych błędów w nich zawartych. Tym samym pogarszają się właściwości predykcyjne modelu.

W celu uniknięcia przeuczenia modelu stosowane są metody wyboru zmiennych istotnych. Ich głównym celem jest identyfikacja zmiennych mających największy wkład do budowy modelu. Model konstruowany dla wybranych zmiennych istotnych ma zbliżone parametry predykcyjne w stosunku do wyjściowego modelu konstruowanego z wykorzystaniem wszystkich zmiennych lecz zazwyczaj mniejszej liczby czynników. Obecnie stosuje się wiele metod wyboru zmiennych wśród których duża część jest

dedykowana metodzie częściowych najmniejszych kwadratów, PLS [36,37]. Są to m.in. metoda zmiennych znaczących dla projekcji (z ang. *variable importance in projection*, VIP) [37], współczynnik selektywności (z ang. *selectivity ratio*, SR) [38,39], metoda eliminacji zmiennych nieistotnych (z ang. *uninformative variable elimination*, UVE) [40], metoda korelacji wieloczynnikowej (z ang. *significance multivariate correlaction*, SMC) [41].

Proponowane metody uwzględniają różnego rodzaju parametry charakteryzujące wkład danej zmiennej do budowy modelu, które są następnie porównywane z wyznaczoną wartością progową tego parametru. Zmienne opisane przez wartości większe niż wartość progowa uznaje się za istotne do konstrukcji modelu, natomiast zmienne o wartościach wyznaczanych parametrów niższych niż wartość progowa są nieistotne.

Metoda zmiennych znaczących dla projekcji, VIP [37] należy do metod wyboru zmiennych bazujących na określonych filtrach pozwalających ocenić istotność poszczególnych zmiennych dla budowy modelu. Stosując metodę VIP dla każdej i-tej zmiennej wyznaczany jest parametr istotności VIP_i, zgodnie z równaniem 7.

$$VIP_i = \sqrt{\frac{\sum_{j=1}^{f} w_{ij}^2 \cdot SSY_j \cdot n}{SSY_t \cdot f}} \tag{7}$$

gdzie, w_{ij} jest wagą dla i-tej zmiennej wyznaczoną na podstawie modelu PLS lub PLS-DA i j-tego czynnika, SSY_j to suma kwadratów wartości zmiennej zależnej **y** uzyskanej dla modelu PLS o j czynnikach, n określa liczbę zmiennych, SSY_t to suma kwadratów wartości zmiennej zależnej **y**, a f to optymalna liczba czynników głównych użyta do konstrukcji modelu.

Zazwyczaj zmienna jest uznawana za istotną do budowy modelu PLS, gdy wartość wyznaczonego dla niej parametru VIP jest większa od 1. Natomiast za umiarkowanie istotną uznaje się zmienną, której parametr VIP znajduje się w przedziale od 0,8 do 1, a mało istotną zmienną charakteryzuje wartość parametru VIP poniżej 0,8 [42]. Procedura eliminacji zmiennych z wykorzystaniem metody VIP może być wykonywana kilkukrotnie w celu zredukowania liczby zmiennych. Przy czym w każdej kolejnej

iteracji konstruuje się nowy model dla zbioru zawierającego zmienne, które zostały w poprzednim kroku uznane za istotne.

Współczynnik selektywności, SR [39] jest kolejną metodą wykorzystywaną do filtrowania zmiennych ze względu na ich istotność do budowy modelu PLS. Podobnie jak w poprzedniej metodzie dla każdej zmiennej wyznaczana jest wartość parametru definiującego jej istotność. W danym podejściu współczynnik selektywności dla każdej zmiennej definiuje się jako stosunek wariancji opisanej przez dany model ($v_{opisana}$) do wariancji reszt od modelu (wariancji reszt od modelu, v_{reszt}) zgodnie z następującym równaniem:

$$SR_i = \frac{v_{i,opisana}}{v_{i,reszt}} \tag{8}$$

gdzie $v_{i,opisana}$ jest wariancją zmiennej i opisaną przez model, $v_{i,reszt}$ jest wariancją reszt od modelu uzyskanych dla i-tej zmiennej .

W metodzie PLS-DA każdy z czynników jest reprezentowany przez wektor wyników o wymiarowości [m×1] i wektor wag o wymiarowości [1×n] uzyskiwany poprzez projekcję macierzy zbioru na znormalizowany wektor wyników. Iloczyn wektora wyników i wektora wag pozwala na uzyskanie macierzy o wymiarowości [m×n], która stanowi projekcję danych na określony czynnik PLS-DA. Tak więc oryginalna macierz zbioru modelowego **X** [m×n] może być przedstawiona jako suma macierzy stanowiącej projekcję danych na określony czynnik PLS i macierzy reszt od modelu o tej samej wymiarowości. Pierwsza z nich zawiera informację opisaną przez model o założonej kompleksowości, natomiast macierz reszt dotyczy informacji nieopisanej przez model. Podczas konstrukcji modelu PLS uwzględniana jest zarówno wariancja danych jak i kowariancja pomiędzy danymi, a zmienną zależną. Parametr SR jest wyznaczany dla każdej zmiennej na podstawie modelu PLS-DA o określonej kompleksowości. Następnie na podstawie założonej wartości granicznej danego parametru zmienne są charakteryzowane jako istotne lub nie. Wartość graniczna SR powyżej której zmienne są uznawane za istotne jest wybierana arbitralnie, w zależności od typu analizowanych danych oraz problemu badawczego.

Metoda korelacji wieloczynnikowej, SMC [41] jest kolejną metodą należącą do grupy metod wyboru zmiennych wykorzystujących filtry. Jej głównym celem jest określenie istotności dla każdej zmiennej na podstawie współczynników regresji skonstruowanego modelu PLS. W metodzie SMC, do odtworzenia macierzy danych na podstawie uzyskanych parametrów modelu PLS wykorzystuje się jako wektor wag. Różnice pomiędzy metodą SMC i SR polegają na pominięciu w metodzie SMC operacji matematycznych warunkujących ortogonalność wariancji i zastosowaniu znormalizowanych wektorów współczynników regresji jako wag do wyznaczania wariancji opisanej przez model.

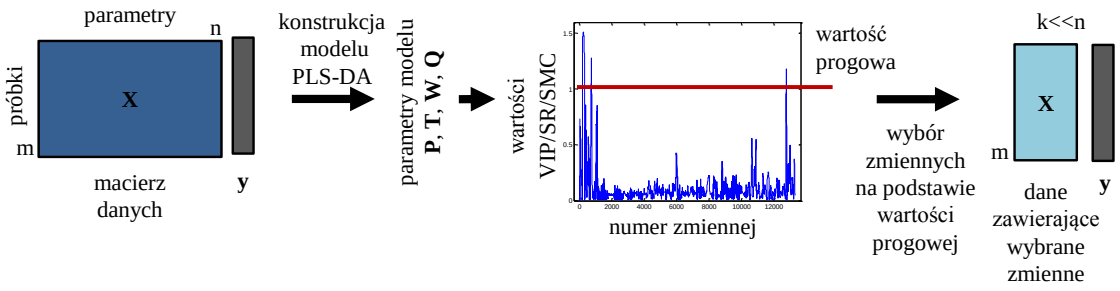
Istotność zmiennej jest określana poprzez współczynnik SMC stanowiący iloraz wariancji opisanej i wariancji reszt wyznaczanych w oparciu o parametry modelu PLS zgodnie z równaniem 9. Zmienne o relatywnie dużej wartości współczynnika SMC wykazują dobrą korelację ze zmienną zależną **y** i tym samym są istotne dla konstrukcji modelu PLS. Wartość graniczna parametru SMC, determinująca istotność zmiennych jest definiowana z użyciem wartości krytycznych dla F-testu na określonym poziomie istotności i danej liczbie stopni swobody zależnej od liczby analizowanych próbek, $F(\alpha,m_1,m_2)$, α stanowi wybrany poziom istotności, $m_1 = 1$, $m_2 = m-2$, a m to liczba próbek.

$$\mathrm{SMC_i} = \frac{\mathrm{v_{SMCi,opisana}}}{\mathrm{v_{SMCi,reszt}}}(\mathrm{m}-2) \tag{9}$$

gdzie $v_{SMCi,opisana}$ jest wariancją zmiennej i opisaną przez model, $v_{SMCi,reszt}$ jest wariancją reszt od modelu uzyskanych dla i-tej zmiennej.

Wszystkie wymienione powyżej metody wyboru zmiennych istotnych wymagają skonstruowania modelu PLS na podstawie którego, wyznaczane są różnego rodzaju parametry charakteryzujące istotność zmiennych do budowy modelu. Następnie ustalana jest wartość progowa danego parametru względem której oszacowuje się istotność danej zmiennej dla budowy modelu. Zmienne o wartościach wyższych niż wartość progowa są uznane za istotne do konstrukcji modelu, natomiast zmienne o wartościach wyznaczanych parametrów niższych niż wartość progowa są uznawane za nieistotne.

Etapy wyboru zmiennych w metodzie PLS z użyciem filtrów VIP, SR i SMC przedstawia Rys. 11.



Rys. 11 Etapy wyboru zmiennych w metodzie częściowych najmniejszych kwadratów z użyciem filtrów VIP, SR i SMC

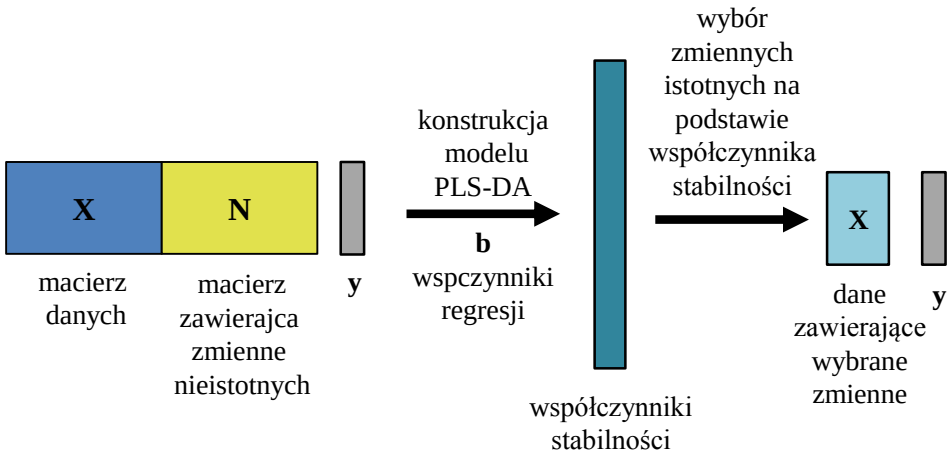
Metoda eliminacji zmiennych nieistotnych, UVE [40] reprezentuje grupę metod wyboru zmiennych tzw. *wrapped methods*. Jako zmienne nieistotne w metodzie UVE-PLS uznawane są te, które posiadają informację porównywalną do informacji zawartej w szumie. Do macierzy danych **X** o wymiarowości [m×n] (m próbek, n zmiennych) dodawana jest macierz zmiennych **N** o bardzo małych wartościach, małej wariancji i znikomej kowariancji ze zmienną **y** o wymiarze [m×n*]. Wynikiem tego jest uzyskanie macierzy **X+N** o m wierszach i n + n* kolumnach (zob. Rys.12). Następnie dla macierzy (**X**+**N**) konstruowany jest model z wykorzystaniem procedury jackknifing, a jego współczynniki regresji są wykorzystywane do wyznaczania współczynników stabilności zgodnie z równaniem:

$$s_i = \frac{\bar{b}_i}{\mathrm{std}(\mathbf{b}_i)} \tag{10}$$

gdzie $\bar{b}_i$ jest średnią wartością współczynników regresji dla i-tej zmiennej i uzyskanych dla modelu PLS-DA z zastosowaniem procedury jackknifing, a std($\mathbf{b}_i$) jest odchyleniem standardowym dla wartości tych współczynników.

Maksymalna bezwzględna wartość współczynnika stabilności dla zmiennych dodanych wyznacza granicę, poniżej której znajdują się zmienne nieistotne. W konsekwencji wybierane są zmienne o stabilności wyższej niż wyznaczona granica i na ich podstawie

konstruowany jest końcowy model PLS. Etapy wyboru zmiennych w metodzie UVE-PLS przedstawia Rys. 12.



Rys. 12 Etapy wyboru zmiennych w metodzie częściowych najmniejszych kwadratów z eliminacją zmiennych nieistotnych

3.3 Przykłady weryfikacji autentyczności wybranych produktów w oparciu o chromatograficzne odciski palca

Chromatograficzne odciski palca, zawierające duży zasób informacji na temat składu chemicznego badanych próbek, są często stosowane do badania autentyczności wybranych produktów, która jest determinowana przez ich skład ilościowy/jakościowy lub ich pochodzenie geograficzne. Monitorowanie jakości produktów ziołowych to jeden z przykładów wykorzystania chromatograficznych odcisków palca. Zioła i ich ekstrakty charakteryzują się złożonym składem, a identyfikacja komponentów w nich zawartych jest niezbędna do oceny ryzyka związanego z ich spożywaniem.

W literaturze techniki bazujące na chromatograficznych odciskach palca są rekomendowane jako rzetelna metodologia do identyfikacji i kontroli jakości leków i preparatów ziołowych [43,44]. Badania tego typu produktów są prowadzone w celu wykrywania zafałszowań spowodowanych przez ich obniżoną jakość. Obecnie jednymi z najbardziej popularnych metod stosowanych do analizy materiałów roślinnych jest analiza chromatograficzna produktów ziołowych bazująca na chromatograficznych odciskach palca w połączeniu z różnego rodzaju narzędziami chemometrycznymi [45,46].

Podobnie jak produkty ziołowe analizuje się także inne produkty farmaceutyczne, których fałszowanie stanowi realne niebezpieczeństwo dla zdrowia i życia ludzi. Fałszowane leki są pozbawione kontroli jakości i tym samym nie można zagwarantować bezpieczeństwa ich stosowania i skuteczności. Obecnie, ocena bezpieczeństwa fałszowanych leków opiera się głównie na identyfikacji i oznaczeniu zawartych w nich substancji czynnych. Jednak coraz częściej dokonuje się również analiz pod kątem obecności potencjalnych toksycznych składników pobocznych powstających w trakcie wytwarzania tego typu leków, takich jak np. pozostałości rozpuszczalników czy innych zanieczyszczeń [2,47].

Kolejnym obszarem wykorzystania chromatograficznych odcisków palca jest badanie produktów spożywczych. Kontrola autentyczności żywności w dużej mierze polega na identyfikacji pochodzenia żywności w związku z koniecznością weryfikacji specyficznych oznaczeń potwierdzających pochodzenie geograficzne produktów. Pochodzenie geograficzne żywności, wraz z jej składem, powinno być znane konsumentowi i oznaczone na każdym produkcie spożywczym. Działania te mają na celu zagwarantowanie bezpieczeństwa, autentyczności produktów oraz ochronę praw

producentów żywności. Ze względu na dużą liczbę produktów wymagających kontroli oraz ciągłe zmiany regulacji prawnych definiujących sposoby oceny autentyczności produktów spożywczych wciąż istnieje duże zapotrzebowanie na opracowanie szybkich i tanich procedur wykrywania zafałszowań produktów i/lub sprawdzających zgodność towaru ze specyfikacją podaną na etykiecie. Dotychczas opracowano wiele metod pozwalających na określenie pochodzenia geograficznego produktów spożywczych [47,48] oraz weryfikację ich autentyczności [49].

W taksonomii zwierząt i roślin pojęcie autentyczności wiąże się z poprawnym przyporządkowaniem zwierząt/roślin do odpowiednich podgrup ewolucyjnych (królestwo, typ, gromada, rząd rodzina, gatunek) na podstawie wykonanych badań dotyczących zawartości poszczególnych substancji w próbkach pochodzenia zwierzęcego/roślinnego. Dziedziną taksonomii opierającą się na badaniu składu chemicznego jest chemotaksonomia, która wykorzystuje informacje o składzie chemicznym w celu ulepszenia systematyki organizmów żywych. Najczęściej wykonywanymi badaniami w taksonomii jest analiza metabolitów pierwotnych i wtórnych oraz nośników informacji genetycznej (kwasy nukleinowe i białka) za pomocą różnego rodzaju technik chromatograficznych [50].

W przemyśle petrochemicznym, ze względu na złożoność analizowanych produktów, analiza chromatograficznych odcisków palca jest bardzo często wykorzystywana do identyfikacji zafałszowania benzyny domieszkowanej rozpuszczalnikami organicznymi [51], zafałszowania oleju napędowego domieszkowanego jadalnym olejem roślinnym [52], czy wykrycia prób nielegalnego usuwania dodatków akcyzowych [53].

Techniki chromatograficzne są także wykorzystywane do analizy próbek pochodzenia środowiskowego pod względem zawartości substancji szkodliwych. Tego typu badania wymagają zastosowania metod analitycznych pozwalających na uzyskanie bardzo niskich granic oznaczalności, które są ściśle zdefiniowane przez odpowiednie normy określające dopuszczalną zawartość substancji szkodliwych w środowisku. Małe stężenia substancji mogą być oznaczane za pomocą metod chromatograficznych sprzężonych z selektywnymi detektorami. Uzyskane w ten sposób chromatograficzne odciski palca posiadają duży zasób informacji ze względu na złożoność składu typowy dla próbek środowiskowych i wymagają zastosowania odpowiednich narzędzi chemometrycznych.

Stale wzrastająca liczba publikacji poświęconych wykorzystaniu sygnałów chromatograficznych w połączeniu z narzędziami chemometrycznymi w kontekście badania autentyczności oraz zgodności produktów z określonymi normami niewątpliwie świadczy o dużym popycie na tego typu rozwiązania.

4. Badania własne

W moich badaniach skupiłam się na wykazaniu przydatności chromatograficznych odcisków palca do weryfikacji autentyczności m.in. oleju napędowego i preparatu Viagra®. Ponadto, badałam możliwość oceny zagrożenia skażenia wody tributylocyną na podstawie chromatograficznych odcisków palca uzyskanych dla próbek wód lądowych i ich dyskryminacji za pomocą metody PLS-DA.

W celu dokonania skutecznej ekstrakcji informacji ze zbioru chromatograficznych odcisków palca wykorzystałam odpowiednio dobrane narzędzia chemometryczne pozwalające wstępnie przygotować dane, a następnie skonstruować oraz rygorystycznie zwalidować (w oparciu o proponowaną przeze mnie metodę walidacji) opracowane modele dyskryminacyjne.

W kolejnych podrozdziałach niniejszej rozprawy doktorskiej przedstawiłam proponowane przeze mnie strategie wspomagające identyfikację zafałszowania paliw z użyciem chromatograficznych odcisków palca rejestrowanych techniką GC-FID, metodę wykorzystania chromatograficznych odcisków palca opisujących profile zanieczyszczeń preparatu Viagra®, rejestrowanych metodą HPLC-DAD do weryfikacji ich autentyczności oraz możliwości zastosowania metod chemometrycznych do analizy chromatograficznych odcisków palca w kontekście oceny ryzyka skażenia wody tributylocyną i usprawnienia funkcjonowania laboratorium.

4.1 Identyfikacja procederu fałszowania oleju napędowego

Na terenie Unii Europejskiej cena paliwa jest uzależniona od czynników ekonomicznych i przepisów prawnych określających wysokość podatku akcyzowego. Na ogół, przyjmuje się różne stawki podatku akcyzowego ze względu na przeznaczenie paliwa. W Polsce na olej napędowy, wykorzystywany do celów grzewczych i napędzania maszyn rolniczych, nałożona jest niższa kwota podatku niż na olej napędowy przeznaczony do regularnego transportu. W celu ułatwienia wizualnego odróżnienia oleju napędowego ze względu na przeznaczenie dodawane są do niego dodatki akcyzowe takie jak czerwony barwnik (Solvent Red 19 lub Solvent Red 164) oraz marker (Solvent Yellow 124) [54]. Stężenia tych komponentów w paliwie są ściśle określone w Rozporządzeniu Ministra Finansów z dnia 20 sierpnia 2010 r. Obecność tych dodatków nie wpływa na właściwości fizykochemiczne paliwa, ani nie ogranicza

jego dalszego przeznaczenia. Znacząca różnica w cenie oleju napędowego stanowi jednak potencjalny impuls do nielegalnej praktyki usuwania dodatków akcyzowych z tańszego paliwa i sprzedawania go po wyższej cenie. Ta procedura jest znana jako tzw. odbarwianie paliwa, gdyż prowadzi do zmiany jego barwy z czerwonej na żółtą, która jest charakterystyczna dla paliwa nie posiadającego dodatków akcyzowych. Odbarwianie paliwa może być realizowane poprzez adsorpcję dodatków fiskalnych wykorzystując powszechnie dostępne materiały lub poprzez zmianę ich struktury. Proceder ten prowadzi do znacznych strat w budżecie państwa i dlatego poszukuje się szybkich metod analitycznych służących do wykrywania zafałszowanego oleju napędowego.

Do tej pory, w ramach prowadzonych przez naszą grupę badań wykazaliśmy możliwość oznaczania zawartości barwnika i znacznika w oleju napędowym oraz wykrywania procederu odbarwiania paliwa na podstawie widm całkowitej fluorescencji [53,55]. W ostatnich badaniach podjęto próbę opracowania nowego podejścia analitycznego w oparciu o standardową technikę badania składu paliwa jaką jest chromatografia gazowa z detekcją płomieniowo-jonizacyjną [56]. Głównym celem przeprowadzonych badań było opracowanie metody pozwalającej na odróżnienie paliwa odbarwianego i nieodbarwianego na podstawie chromatograficznych odcisków palca analizowanych próbek. W celu rozwiązania podjętego problemu badawczego w pierwszym etapie badań wykonano eksperyment polegający na symulacji procederu odbarwiania oleju napędowego za pomocą adsorpcji dodatków akcyzowych. Chromatograficzne odciski palca były rejestrowane dla próbek paliwa przed i po procesie odbarwiania stosując chromatografię gazową z detekcją płomieniowo-jonizacyjną. Należy podkreślić, iż ta technika nie pozwala analizować zmian zawartości znacznika i barwnika, ponieważ w temperaturze w jakiej prowadzony jest rozdział substancje te ulegają rozkładowi. Proponowana metoda nie może opierać się na analizie ilościowej znacznika i barwnika, gdyż te z założenia zostają usunięte w procesie odbarwiania. Z tego powodu, zaproponowałam aby skupić się na profilach chromatograficznych stanowiących chemiczne odciski palca badanych próbek, które pozwolą w sposób całościowy opisać różnice w ogólnym składzie próbek spowodowane procesem odbarwiania. Takie podejście wymaga analizy fluktuacji składu chemicznego paliwa wywołanych jego odbarwianiem.

Eksperyment polegający na symulacji procesu odbarwiania oleju napędowego wykonano w Laboratorium Izby Celnej w Białej Podlaskiej. Analizie poddano łącznie 31 próbek oleju napędowego. Pochodziły one od różnych polskich dostawców i zostały pobrane zgodnie z wymogami określonymi w normie PN-EN ISO 3170: 2004. Chromatograficzne odciski palca zarejestrowano za pomocą chromatografu typu Agilent Technologies 6890N wyposażonego w detektor płomieniowo-jonizacyjny (GC-FID) i kolumnę Restek (60 m×0,25 mm). Jako gaz nośny zastosowano hel o czystości 5,0 i przepływie 1,3 $mL \cdot min^{-1}$. Objetość pojedynczego nastrzyku próbki wynosiła 0,1µL. Natomiast temperatura analizy wzrastała o 3 °C / min w granicach od 50 °C do 320 °C. Całkowity czas analizy wynosił 100 min.

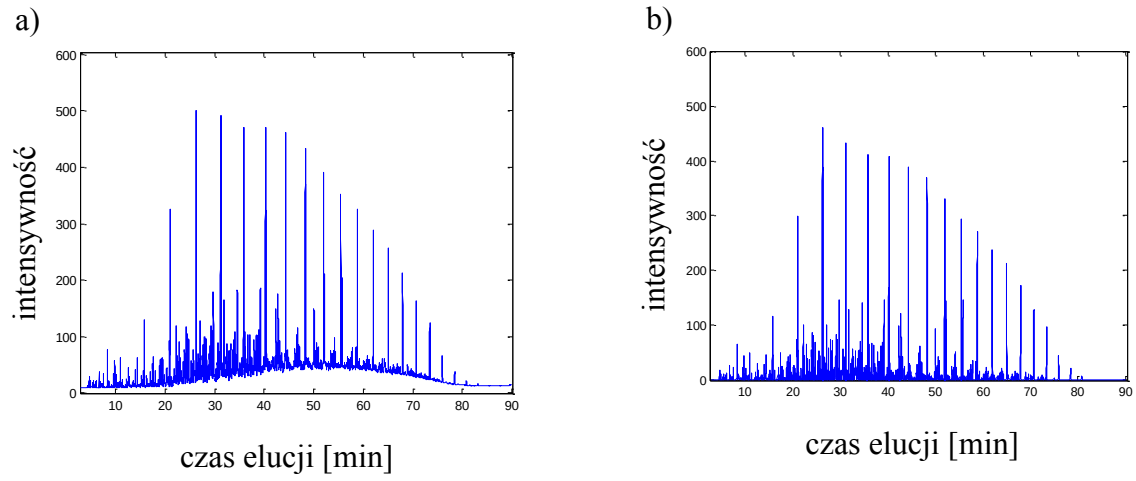
Dla uzyskanego zbioru chromatogramów zmniejszono częstość próbkowania poprzez redukcję liczby punktów pomiarowych stosując liniową interpolację. Początkowo, chromatogramy zawierały 104 399 punktów pomiarowych. Zabieg ten zastosowano w celu przyspieszenia prowadzenia dalszych obliczeń oraz lepszego działania stosowanych algorytmów. Analizowano kilka wariantów częstości próbkowania prowadzących do uzyskania chromatogramów zawierających od 5 000 do 50 000 punktów pomiarowych. Jako optymalną liczbę wybrano 25 000 punktów pomiarowych, które pozwoliły zachować oryginalny kształt wyjściowych sygnałów chromatograficznych.

Na podstawie wizualnej oceny zbioru chromatogramów stwierdzono, że wymagają one usunięcia linii podstawowej i skorygowania przesunięć pików. Do eliminacji linii podstawowej zastosowano algorytm asymetrycznych najmniejszych kwadratów z funkcją kary (z ang. *penalized asymmetric least squares*, PAsLS) [13]. Przeanalizowano wyniki uzyskane dla różnych wartości parametrów wyjściowych. Najlepszy kształt linii podstawowej uzyskano stosując drugą pochodną i następujące wartości parametrów $\lambda = 10^4$ i $p = 10^{-3}$.

Następnie, skorygowano przesunięcia pików techniką maksymalizującą wzajemną korelację sygnałów (z ang. *correlation optimized warping*, COW). Wszystkie chromatogramy nałożono względem sygnału wzorcowego wybranego zgodnie z metodą opisaną w [15].

Przeanalizowano wiele zestawów parametrów wejściowych dla metody COW. Najlepsze rezultaty nakładania sygnałów zostały uzyskano w przypadku gdy sygnały

chromatograficzne GC-FID były podzielone na 250 sekcji, a parametr elastyczności był równy 3. Przykładowy chromatogram próbki oleju napędowego przed i po wstępnym przygotowaniu danych przedstawia Rys. 13.



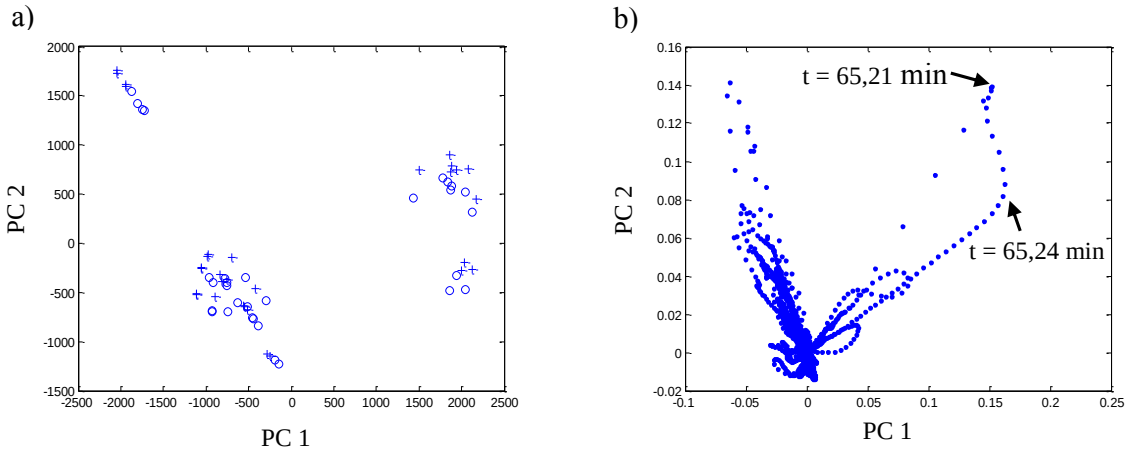
Rys. 13 Przykładowy chromatogram GC-FID próbki oleju napędowego (a) przed i (b) po usunięciu linii podstawowej za pomocą metody asymetrycznych najmniejszych kwadratów z funkcją kary i nałożeniu chromatogramów z wykorzystaniem metody zoptymalizowanego nakładania sygnałów maksymalizującej ich wzajemną korelację

Chromatograficzne odciski palca po uprzedniej zmianie częstości próbkowania (ze 104 399 do 25 000 punktów pomiarowych), usunięciu linii podstawowej ($\lambda = 10^4$ i $p = 10^{-3}$) i nałożeniu pików (N = 250, s = 3) zestawiono w macierz danych o wymiarach [62×25 000].

Eksplorację sygnałów GC-FID przeprowadzono za pomocą analizy czynników głównych. Efektywna kompresja danych wskazuje na dużą liczbę skorelowanych zmiennych, gdyż pierwsze dwa czynniki główne opisują aż 73,7% całkowitej wariancji danych. Projekcja próbek na przestrzeń zdefiniowaną przez pierwsze dwa czynniki główne umożliwia ocenę ich podobieństwa chemicznego. Każdy punkt na projekcji PC 1-PC 2 (zob. Rys. 14) to pojedynczy chromatogram GC-FID charakteryzujący daną próbkę. Analizując położenie próbek wzdłuż osi PC 1 można zaobserwować dwie dobrze oddzielone grupy próbek oleju napędowego. Jednak, na taki wynik grupowania nie ma wpływu przeprowadzony proces odbarwiania oleju napędowego.

Analiza projekcji parametrów opisanych przez pierwsze dwa czynniki główne pozwoliła zidentyfikować zakresy czasów elucji odpowiadające substancjom odpowiedzialnym za

grupowanie się próbek wzdłuż osi PC 1. Są to dwa piki pochodzące od substancji, którym odpowiadają czasy elucji około 65,21 min. i 65,24 min. odpowiadające estrom metylowym kwasów tłuszczowych (FAME). Związki FAME nie mogą być uznane jako potencjalne markery wykrywania procederu fałszowania oleju napędowego, ponieważ są celowo dodawane do paliwa w trakcie jego produkcji. W Polsce dopuszcza się obecność FAME w oleju napędowym w ilości do 7% (v/v). Różnice w ich zawartości w różnych partiach oleju napędowego mogą być znaczne i najczęściej zależą od producenta danego paliwa. Grupa próbek oleju napędowego charakteryzowana przez pozytywne wartości PC 1 zawiera FAME, w ilości od 4,1% do 5,3 (v/v). Grupowanie się próbek ze względu na ich odbarwienie także nie zostało zaobserwowane dla projekcji opisanych przez inne pary czynników głównych. Fakt, że analiza eksploracyjna z wykorzystaniem metody PCA nie pozwala rozróżnić próbek oleju napędowego ze względu na jego zafałszowanie nie wyklucza potencjalnej możliwości ich dyskryminacji stosując metody uczenia z nadzorem np. analizę dyskryminacyjną.



Rys. 14 Projekcja próbek oleju napędowego na płaszczyznę zdefiniowaną przez pierwszy i drugi czynnik główny. Próbki oryginalne oznaczono jako (○), a próbki odbarwione jako (+)

W celu rozróżnienia odbarwionych i nieodbarwionych próbek oleju napędowego na podstawie chromatograficznych odcisków palca wykorzystano metodę PLS-DA. Model dyskryminacyjny skonstruowano dla zbioru modelowego złożonego z 21 chromatogramów oryginalnych próbek oleju napędowego wybranych za pomocą algorytmu Kennarda i Stonea oraz 21 chromatogramów uzyskanych dla tych samych próbek po ich odbarwieniu. Pozostałe 20 chromatogramów próbek paliwa przed i po

procesie odbarwienia utworzyło zbiór testowy użyty w celu określenia właściwości predykcyjnych modelu dyskryminacyjnego. Optymalny model PLS-DA miał kompleksowość równą 6, którą ustalono za pomocą walidacji krzyżowej typu ‘wyrzuć jedną próbkę’. Utworzone reguły logiczne pozwoliły poprawnie rozpoznać wszystkie próbki paliwa zarówno ze zbioru testowego jak i ze zbioru modelowego (zob. Tabela 1). W celu określenia zdolności predykcyjnych modelu wyznaczano średnie wartości parametrów walidacyjnych z wykorzystaniem podejścia ‘bootstrap’. Każdorazowo ze zbioru modelowego losowano ze zwracaniem 21 próbek nieodbarwionych, których chromatogramy przed i po odbarwieniu były wykorzystywane do konstrukcji modelu dyskryminacyjnego. Proces ten powtarzano 1000 razy otrzymując 1000 zestawów parametrów walidacyjnych. Uzyskane rezultaty analizy dyskryminacyjnej wspierają hipotezę, iż na podstawie sygnałów chromatograficznych GC-FID próbek oleju napędowego można rozróżnić próbki odbarwione i nieodbarwione.

W kolejnym etapie badań, zastosowano różne metody wyboru zmiennych istotnych w celu uniknięcia możliwości przeuczenia modelu dyskryminacyjnego i wyznaczenia istotnych regionów czasu elucji, w których ulegały wymyciu związki istotne w kontekście dyskryminacji badanych próbek oleju napędowego. W metodzie UVE-PLS-DA do każdego z 1000 wylosowanych zbiorów modelowych dodawano macierz danych o takiej samej liczbie wierszy co macierz stanowiąca zbiór modelowy i liczbie kolumn równej 10 000. Macierz ta zawierała zmienne nieistotne, którymi były liczby wybrane losowo z rozkładu normalnego i pomnożone przez współczynnik $c = 10^{-12}$. Następnie, w każdej iteracji konstruowano model PLS-DA, a jako istotne zmienne wybierano te, których wartość bezwzględna stabilności, liczona na podstawie uzyskanych współczynników regresji, była większa niż maksymalna wartość stabilności zmiennych nieistotnych.

Ten sam sposób wielokrotnego losowania ze zwracaniem zastosowano w pozostałych metodach wyboru zmiennych. W metodach VIP i SR w pierwszym kroku konstruowano model PLS-DA o założonej kompleksowości, a następnie dla każdej zmiennej wyznaczano dla poszczególnej metody odpowiedni parametr istotności. Stosując podejście ‘bootstrap’, dla każdej zmiennej uzyskano 1000 wartości zarówno parametru VIP jak i parametru SR. Jako istotne wybierane były te zmienne, których wartości średnie parametrów przekraczały odpowiednie wartości progowe. Dla każdej z zastosowanych metod przeanalizowano różne wartości progowe uznając za optymalne

wartości 0,8 dla metody VIP i 1,0 dla metody SR. Współczynnik selektywności obliczany był raz dla wszystkich zmiennych natomiast w metodzie VIP zastosowano trzy iteracje, w których w kolejnej iteracji wykorzystywano zmienne uznane za istotne w poprzednim kroku. Wszystkie metody wyboru zmiennych zaimplementowane w trakcie konstrukcji dyskryminacyjnego wariantu metody częściowych najmniejszych kwadratów, UVE-PLS-DA, SR-PLS-DA i VIP-PLS-DA wykazują ich potencjalną możliwość do wykrywania nielegalnego procederu odbarwiania paliw. Świadczą o tym wyniki otrzymane dla każdej z zastosowanych metod dyskryminacji wraz z parametrami opisującymi modele dyskryminacyjne, ujęte w Tabeli 1.

Tabela 1 Wyniki modeli PLS-DA bez i z zastosowaniem metod wyboru zmiennych (wartości wyrażone w %) (UVE - metoda eliminacji zmiennych nieistotnych, VIP - metoda zmiennych znaczących dla projekcji, SR - współczynnik selektywności, n - liczba zmiennych, f - liczba czynników głównych, SE - czułość, SP - specyficzność, AUC - parametr AUC)

Model	n	f	Zbiór modelowy			Zbiór testowy		
			SE	SP	AUC	SE	SP	AUC
PLS-DA	25 000	6	100,00	100,00	100,00	100,00	100,00	100,00
UVE	14	9	100,00	100,00	100,00	100,00	100,00	100,00
VIP	265	6	90,00	99,60	99,60	90,00	99,60	97,00
SR	16	3	100,00	100,00	100,00	100,00	100,00	100,00

Najgorsze parametry predykcyjne uzyskano dla modelu skonstruowanego za pomocą metody VIP-PLS-DA, wykorzystując do tego największa liczbę zmiennych istotnych. Pozostałe dwie metody wykazały czułość i specyficzność równe 100% dla zbioru testowego dla niewielkiej liczby zmiennych (14 i 16 zmiennych). Dokładniejsza analiza wybranych zmiennych wykazała, że zmienne wybrane z wykorzystaniem metody

SR-PLS-DA odpowiadają czasom elucji substancji polarnych. Obniżanie się stężeń danych substancji w procesie odbarwiania paliwa może być spowodowane przez adsorpcje związków polarnych na złożu, na którym są także sorbowane dodatki akcyzowe paliwa. Dlatego wydaje się, że metoda PLS-DA wraz z metodą wyboru zmiennych opartą na współczynniku selektywności jest najlepszą metodą do wykrywania nielegalnego procederu odbarwienia oleju napędowego. W Tabelach 2 i 3 wyszczególniono związki odpowiadające czasom elucji wybranym przez metody UVE-PLS-DA i SR-PLS-DA.

Przebadanie większej liczby próbek oleju napędowego pozwoliłoby uwzględnić w konstrukcji modelu dyskryminacyjnego większą zmienność danych, co potencjalnie spowodowałoby zwiększenie poprawności przewidywania przynależności nowych próbek paliwa do odpowiednich grup. Jednakże uzyskane wyniki wyraźnie wykazują duży potencjał stosowanej metody, która może wspomóc wymiar sprawiedliwości w walce z nielegalnym usuwaniem dodatków akcyzowych z oleju napędowego.

Tabela 2 Związki chemiczne zidentyfikowane przy użyciu metody UVE-PLS-DA wspierające diagnostykę procesu odbarwiania oleju napędowego

Numer piku	Czas elucji [min]	Zidentyfikowany związek
1	23,937 23,940 23,944	1-metylo-3-propylobenzen ($C_{10}H_{14}$)
2	24,786 24,789 24,793 24,796	1-metylo-4-propylobenzen ($C_{10}H_{14}$)
3	25,398 25,402	1-etylo-2,4-dimetylobenzen ($C_{10}H_{14}$)
4	27,556 27,559 27,563	1,2,3,5-tetrametylobenzen ($C_{10}H_{14}$)
5	40,881 40,884	n-parafiny (C_{14})

Tabela 3 Związki chemiczne zidentyfikowane przy użyciu metody SR-PLS-DA wspierające diagnostykę procesu odbarwiania oleju napędowego (NI - związek niezidentyfikowany)

Numer piku	Czas elucji [min]	Zidentyfikowany związek
1	7,052 7,055	NI
2	23,982 23,985 23,989 23,991 23,996 23,999	4-etyloheptan (C_9H_{20}) lub 1-oktano-2-butyl ($C_{12}H_{26}$)
3	32,292 32,296 32,299	fitol ($C_{20}H_{40}O$)
4	32,891 32,894 32,898	związki zawierające tlen np. 1-propeno-2-nitro-3- (1-cyklooktanyl) ($C_{11}H_{17}NO_2$)
5	38,508	NI
6	47,058	3-metylopentadekan ($C_{16}H_{34}$)

Więcej szczegółów dotyczących identyfikacji procederu fałszowania oleju napędowego z wykorzystaniem narzędzi chemometrycznych przedstawiłam w publikacji „Detection of discoloration in diesel fuel based on gas chromatographic fingerprints", Analytical and Bioanalytical Chemistry, 407 (2015) 1159-1170, która stanowi Załącznik nr 1 do niniejszej rozprawy doktorskiej.

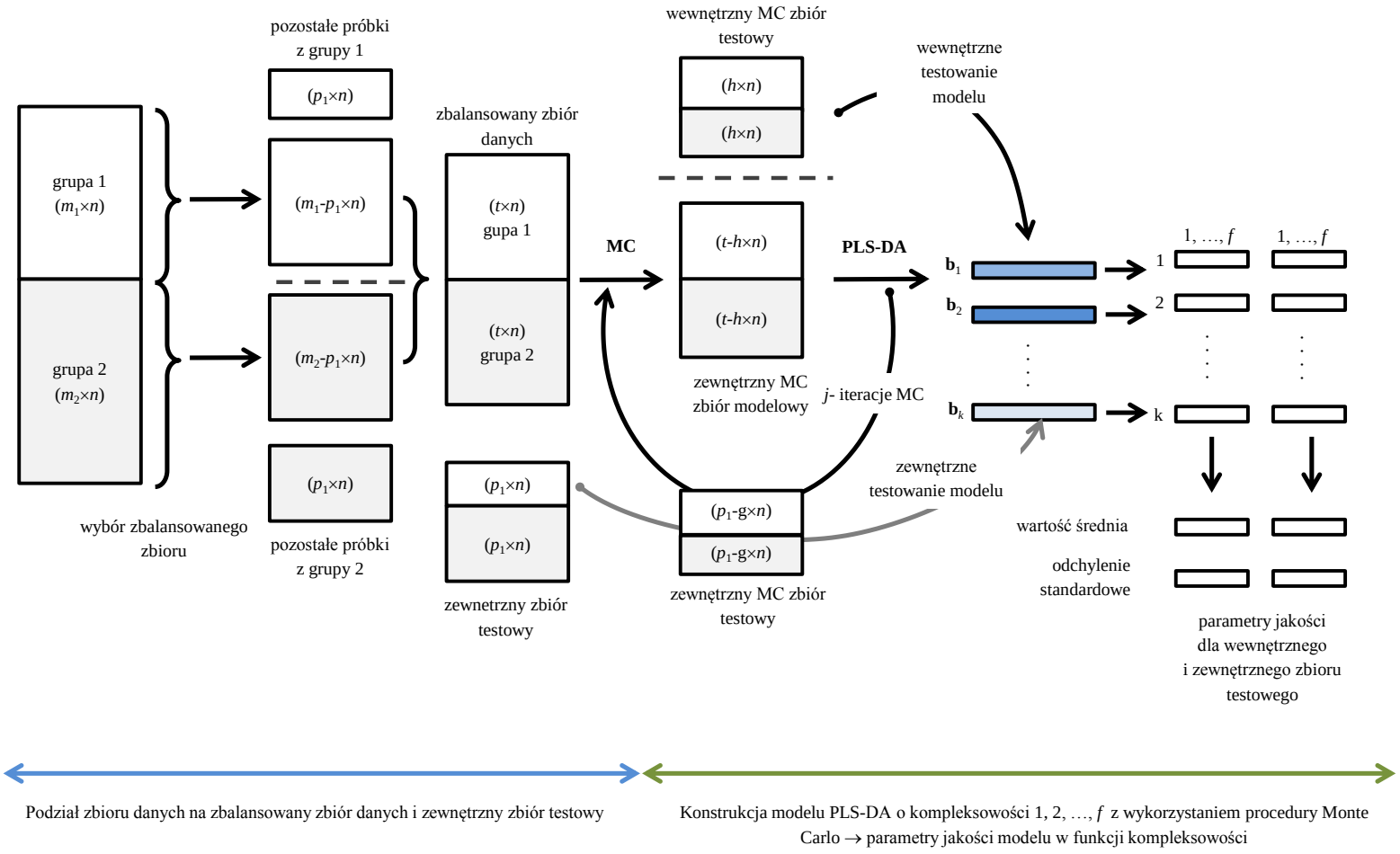
4.2 Nowa metoda walidacji modeli dyskryminacyjnych

Jeden z etapów moich badań obejmował zaproponowanie nowego podejścia do walidacji modelu PLS-DA. W przypadku modeli dyskryminacyjnych konstruowanych dla sygnałów instrumentalnych o dużej liczbie zmiennych istnieje ryzyko tzw. „przeuczenia modelu". Zjawisko to może prowadzić do uzyskania bardzo dobrych wyników dla próbek ze zbioru modelowego, natomiast znacznie gorszych dla próbek zbioru testowego [57]. W celu zniwelowania ryzyka przeuczenia modelu należy oszacować optymalną kompleksowość modelu i/lub zredukować liczbę zmiennych poprzez zastosowanie metod wyboru zmiennych istotnych do jego konstrukcji. Dyskryminacyjny wariant metody częściowych najmniejszych kwadratów jest jedną z najczęściej stosowanych technik dyskryminacyjnych, która pozwala odróżnić próbki należące do różnych grup. Konstrukcja modelu PLS-DA polega na budowaniu reguł logicznych w taki sposób, aby maksymalizować opis wariancji danych z jednoczesną maksymalizacją kowariancji pomiędzy macierzą danych, a zmienną zależną **y**, która opisuje przynależność próbek do odpowiednich grup a zbiorem danych. Poprawność przewidywania modeli dyskryminacyjnych jest określana na podstawie parametrów walidacyjnych obliczanych z wykorzystaniem miedzy innymi różnych metod kroswalidacji np. typu wyrzuć jedną próbkę, metod przepróbkowania danych ('bootstrap', jakknifing, Monte Carlo) oraz wykorzystując niezależny zbiór testowy. Poprawna walidacja modelu dyskryminacyjnego zakłada zastosowanie niezależnego zbioru testowego, czyli zbioru próbek, które nie były wykorzystywane do konstrukcji modelu za pomocą, których oceniane są jego właściwości predykcyjne. Poprawność działania modeli zarówno dyskryminacyjnych jak i klasyfikacyjnych jest uwarunkowana reprezentatywnością zbioru modelowego na podstawie, którego dany model jest konstruowany. W celu uzyskania zbioru modelowego jak najlepiej opisującego wariancję całego zbioru danych wykorzystuje się takie narzędzia jak algorytm Duplex czy metoda Kennarda i Stonea (rozdz. 3.2.2). Wybrany zbiór modelowy służy do konstrukcji modelu dyskryminacyjnego, którego kompleksowość jest zazwyczaj estymowana poprzez zastosowanie różnych wariantów kroswalidacji [33]. Właściwości predykcyjne każdego modelu dyskryminacyjnego są wyznaczane za pomocą takich parametrów jak procent poprawnej klasyfikacji CCR, czułość SE oraz dokładność SP. Parametry te są wyznaczane zarówno dla zbioru modelowego jak i dla zboru testowego.

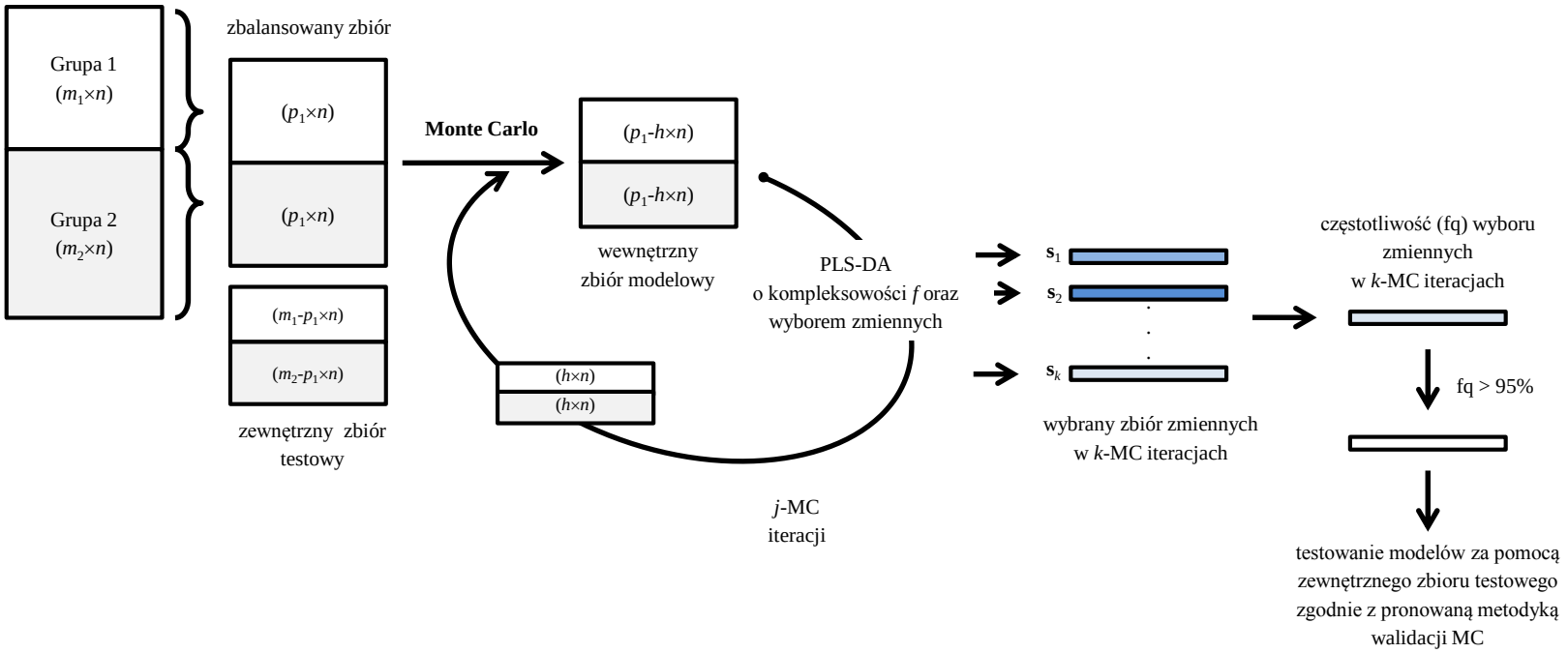
Proponowane podejście walidacyjne obejmuje wykorzystanie procedury Monte Carlo do dwuklasowego problemu dyskryminacyjnego analizowanego za pomocą dyskryminacyjnego wariantu metody częściowych najmniejszych kwadratów. Dodatkowo, metodyka walidacyjna została wykorzystana do ewaluacji modeli PLS-DA umożliwiających wybór zmiennych.

W pierwszym kroku proponowanej procedury z danych losowo wyodrębniany jest zbalansowany podzbiór danych. Pozostałe obiekty stanowią zbiór testowy. Następnie zbalansowany podzbiór danych jest dzielony na zbiór modelowy i wewnętrzny zbiór testowy. Podział ten wykonywany jest wielokrotnie, a w każdej pojedynczej iteracji zarówno zbiór modelowy jak i wewnętrzny zbiór testowy charakteryzują się stałą liczebnością. Ze zbioru testowego wielokrotnie losowany jest zewnętrzny zbiór testowy służący do niezależnej walidacji konstruowanego modelu. Zarówno zbiór modelowy jak i oba zbiory testowe są zbalansowane i wydzielane ze zbioru danych za pomocą procedury Monte Carlo. W każdej pojedynczej iteracji, na podstawie wybranego zbioru modelowego, konstruowany jest model dyskryminacyjny PLS-DA o określonej kompleksowości (1, 2, 3, …, f). Następnie, model jest walidowany za pomocą wewnętrznego i zewnętrznego zbioru testowego poprzez wyznaczenie różnego rodzaju parametrów walidacyjnych. Zewnętrzny zbiór testowy charakteryzuje się tym, że jest on w pełni niezależny ponieważ obiekty, które wchodzą w jego skład nigdy nie biorą udziału w konstrukcji reguł dyskryminacyjnych. Natomiast wielokrotny losowy podział zbalansowanego podzbioru danych na zbiór modelowy i wewnętrzny zbiór testowy dopuszcza sytuacje, w której ten sam obiekt w różnych iteracjach będzie raz w zbiorze modelowym, a raz w zbiorze testowym. Należy zaznaczyć, że proponowane podejście nie dopuszcza w pojedynczej iteracji testowania modelu za pomocą próbek, które były wykorzystane do jego budowy. Wykonując wielokrotnie całą procedurę konstrukcji modelu PLS-DA dla różnej liczby czynników PLS uzyskamy rozkład parametrów walidacyjnych w funkcji kompleksowości modelu. Każdy parametr walidacyjny opisuje jego wartość średnia i odchylenie standardowe wszystkich uzyskanych wyników dla modeli o tej samej kompleksowości. Zaproponowana metoda walidacji modelu PLS-DA pozwala bezpośrednio określić jego optymalną kompleksowość z jednoczesną estymacją wartości parametrów walidacyjnych. Schemat omówionego podejścia walidacji modelu PLS-DA przedstawia Rys. 15.

Procedura Monte Carlo może być także stosowana do walidacji wyboru zmiennych wykorzystując metody omówione w paragrafie 3.2.2. Zaimplementowanie proponowanego podejścia pozwala estymować parametry opisujące istotność poszczególnych zmiennych do budowy modelu dyskryminacyjnego wraz z wyznaczeniem częstości z jaką zmienna była uznana za istotną z wykorzystaniem procedury Monte Carlo (Rys. 16).

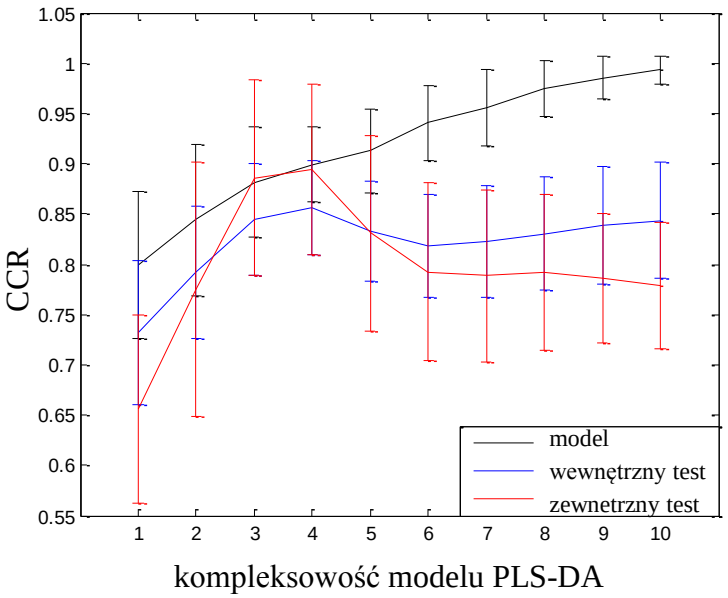


Rys. 15 Ogólny schemat walidacji dyskryminacyjnego wariantu metody częściowych najmniejszych kwadratów z wykorzystaniem procedury Monte Carlo



Rys. 16 Ogólny schemat procedury wyboru zmiennych połączony z dyskryminacyjnym wariantem metody częściowych najmniejszych kwadratów i metodą Monte Carlo pozwalający na określenie częstotliwości wyboru zmiennych istotnych

Końcowy zbiór modelowy zawiera zmienne uwzględniając uprzednio założoną częstotliwość wyboru. Oznacza to, że do konstrukcji modelu wykorzystuje się jedynie te zmienne, które były uznawane za istotne w określonej liczbie iteracji (np. zmienne, które są uznawane za istotne w 95% powtórzeń). Model skonstruowany na podstawie wyznaczonych zmiennych istotnych jest następnie walidowany zgodnie z procedurą przedstawioną na Rys. 15 za pomocą wewnętrznego i zewnętrznego zbioru testowego, a oba zbiory testowe zawierają te same zmienne istotne co zbiór modelowy. Główną zaletą proponowanej metody walidacji jest możliwość estymacji parametrów charakteryzujących jakość konstruowanych modeli dla wielu kompleksowości jednocześnie, co przedstawiono na przykładzie zależności wartości średnich współczynnika poprawnej klasyfikacji od kompleksowości modelu PLS-DA (Rys. 17). Pozwala to dokładniej ocenić jaka liczba ukrytych zmiennych będzie optymalna dla konstrukcji modelu. Proponowane podejście walidacyjne może być zaimplementowane zarówno do problemów dyskryminacyjnych jak i klasyfikacyjnych obejmujących wiele obszarów badań m.in. badania autentyczności leków, produktów spożywczych czy badań dotyczących oceny zgodności składu próbki z wymaganymi normami.



Rys. 17 Wykres zależności wartości średnich współczynnika poprawnej klasyfikacji (CCR) od kompleksowości modelu skonstruowanego z wykorzystaniem dyskryminacyjnego wariantu metody częściowych najmniejszych kwadratów z wartościami odchyleń standardowych wyznaczonych na podstawie procedury Monte Carlo (1000 iteracji)

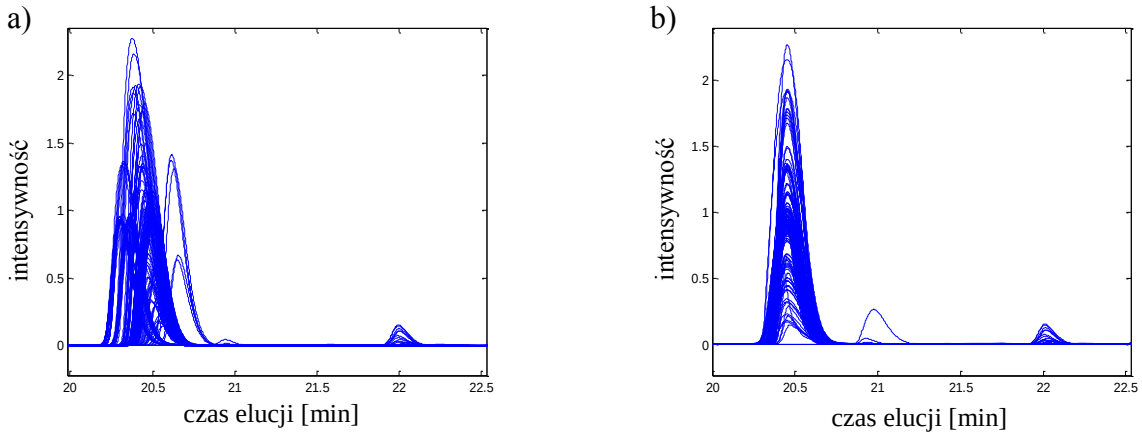
Więcej szczegółów dotyczących nowej metody walidacji modeli dyskryminacyjnych znajduje się w publikacji „The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles", Analyst, 141 (2016) 1060-1070, która stanowi Załącznik nr 2 do niniejszej rozprawy doktorskiej.

4.3 Identyfikacja zafałszowań leku Viagra®

Obecność na rynku zafałszowanych leków to problem niosący poważne zagrożenie dla zdrowia publicznego. Znaczący wzrost w ostatnich latach dostępności na rynku leków, które nie spełniają wymogów jakości może być związany z łatwiejszym dostępem fałszerzy do technologii umożliwiających „kopiowanie" składu oryginalnych leków [58]. W celu badania zgodności wytwarzanych leków z określonymi normami zazwyczaj stosuje się wiele technik instrumentalnych [59]. Autentyczność produktów farmaceutycznych może być analizowana za pomocą różnych technik instrumentalnych takich jak na przykład spektroskopia bliskiej podczerwieni, NIR (z ang. *near infrared spectroscopy*) [60], fluorescencja rentgenowska z dyspersją energii, ED-XRF (z ang. *energy-dispersive X-ray spectroscopy*) [61], spektroskopia magnetycznego rezonansu jądrowego, NMR (z ang. *nuclear magnetic resonance*) [62] czy spektroskopia Ramana (*z ang. Raman spectroscopy*) [63]. Dla leków w postaci tabletek ich autentyczność może również być potwierdzana przez badanie pewnych parametrów fizycznych, takich jak np. grubość tabletki, jej długość oraz masa, jak również poprzez analizę porównawczą zdjęć tabletek wykonanych w tych samych warunkach [64].

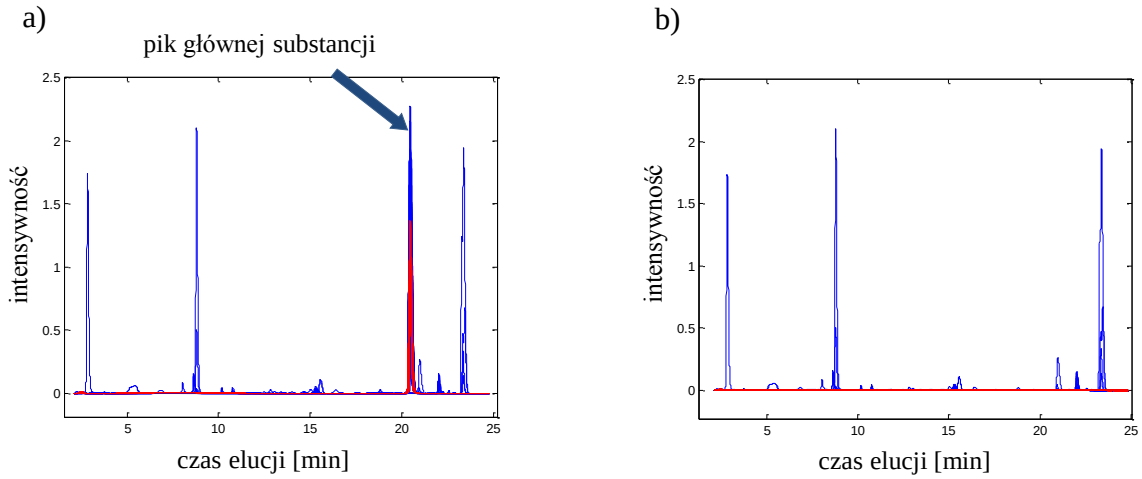
Głównym celem przeprowadzonych badań było wykazanie użyteczności nowego podejścia walidacji wieloparametrowych modeli dyskryminacyjnych, zastosowanych w celu weryfikacji autentyczności próbek leku Viagra® na podstawie ich chromatograficznych odcisków palca [65]. Eksperyment obejmujący analizę składu leku z wykorzystaniem wysokosprawnej chromatografii cieczowej z matrycą diodową, HPLC-DAD został przeprowadzony w Instytucie Zdrowia Publicznego w Brukseli. Przeanalizowano 46 oryginalnych i 97 zafałszowanych próbek leku Viagra®. Otrzymane sygnały chromatograficzne, po ich wstępnym przygotowaniu, posłużyły do konstrukcji wieloparametrowych modeli z wykorzystaniem dyskryminacyjnego wariantu metody częściowych najmniejszych kwadratów. Ponieważ chromatograficzne odciski palca miały jednakową liczbę punktów pomiarowych (13 620), nie wymagały one ponownego przepróbkowania. Z powodu różnic występujących w intensywności linii podstawowej analizowanych sygnałów do jej usunięcia zastosowano metodę częściowych najmniejszych kwadratów, PAsLS. Testując różne wartości parametrów, jako najbardziej odpowiednie uznano $\lambda = 10^5$ oraz $p = 10^{-3}$. Kolejnym krokiem wstępnego przygotowania danych była eliminacja przesunięć pików chromatograficznych. W tym celu zastosowano metodę COW. Spośród wszystkich przeanalizowanych kombinacji wartości parametru elastyczności

s i liczby sekcji najlepsze wyniki uzyskano, gdy sygnały były podzielone na 28 części (27 pierwszych sekcji zawierało po 500 punktów pomiarowych, natomiast ostatnia sekcja składała się ze 120 punktów pomiarowych), a parametr elastyczności wynosił 3. Dodatkowo, skorygowano przesunięcia pików w zakresach czasów elucji od 10,10 min. do 10,43 min. oraz od 20,01 min. do 21,05 min. z liczbami sekcji równymi odpowiednio 10 i 20 oraz parametrami elastyczności równymi 3 i 6. Na Rys. 18 przedstawiono fragment chromatograficznych odcisków palca analizowanych próbek zawierający pik pochodzący od głównej substancji przed i po wstępnym przygotowaniu danych.



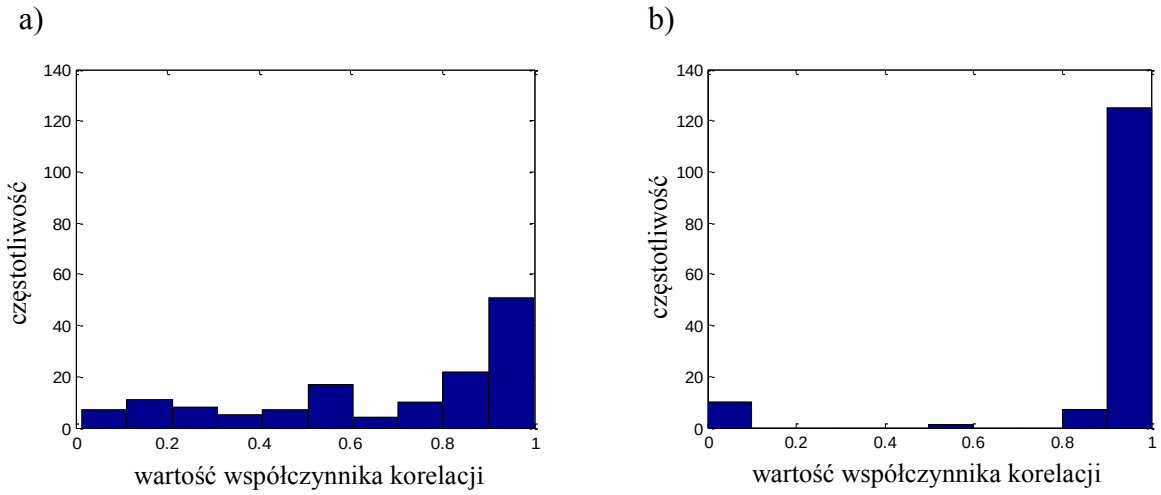
Rys. 18 Fragment chromatogramów zawierający pik pochodzący od Sildenafilu (substancji aktywnej leku Viagra®): (a) przed i (b) po wstępnym przygotowaniu danych

Do dalszej analizy chemometrycznej wykorzystano jedynie profile zanieczyszczeń próbek leku Viagra®, tj. sygnały chromatograficzne otrzymane po usunięciu z oryginalnych sygnałów piku głównego składnika leku, Sildenafilu, który występuje przy czasie elucji ok. 22,5 min. Tym samym, prowadzona analiza była ukierunkowana na wyznaczenie pików odpowiadających substancjom różnicującym próbki autentyczne i zafałszowane, które mogą stanowić markery nielegalnego procederu fałszowania badanego leku. Chromatogramy przed i po usunięciu piku głównego składnika przedstawia Rys. 19.



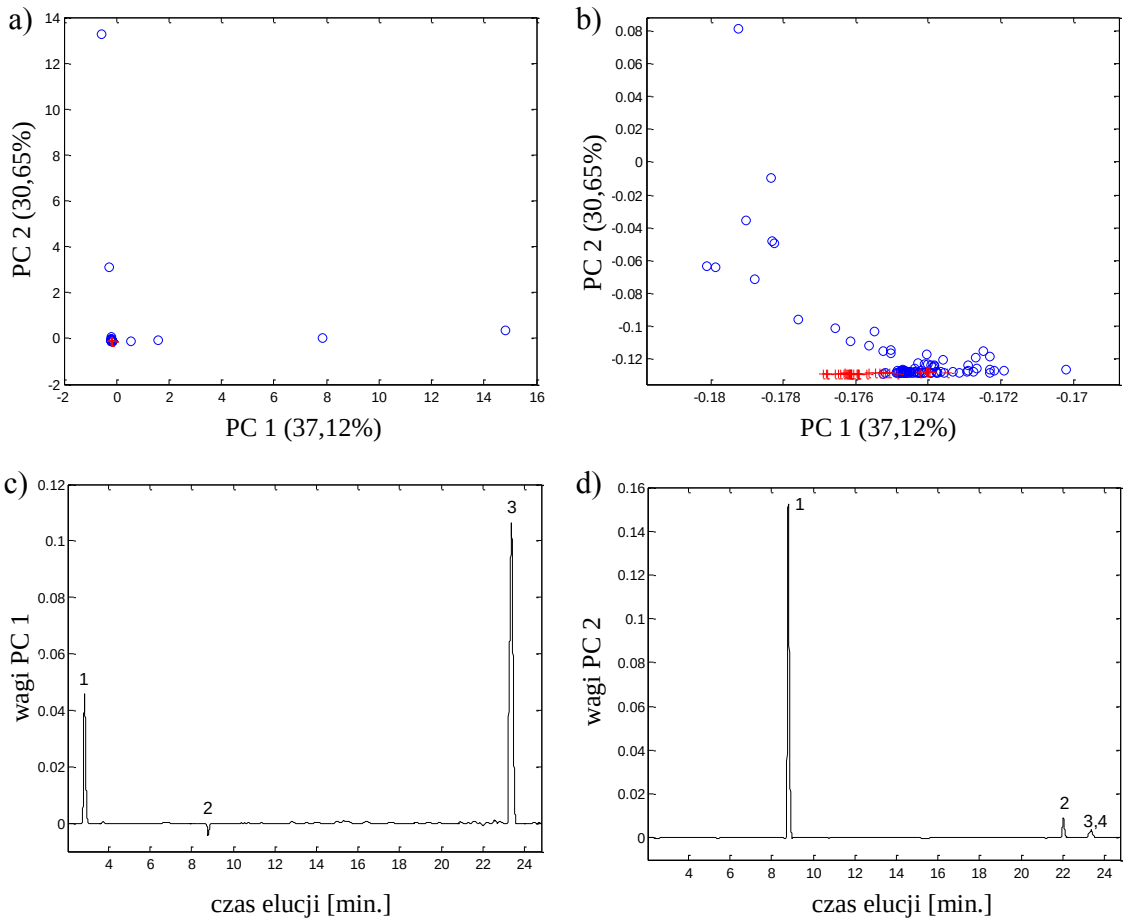
Rys. 19 Chromatogramy HPLC-DAD leku Viagra® zarejestrowane dla długości fali 254 nm przed (a) i po (b) usunięciu piku głównego składnika Sildenafilu

Początkowe wartości współczynników korelacji pomiędzy profilami zanieczyszczeń, a sygnałem względem, którego były one nakładane zawierały się w przedziale od 0,0134 do 0,9988. Natomiast po usunięciu linii podstawowej i korekcji przesunięć pików 88% próbek miało wartość współczynnika korelacji powyżej 0,9. Część sygnałów chromatograficznych uzyskanych dla leku Viagra® pomimo wstępnego przygotowania danych nadal wykazywały małe wartości współczynnika korelacji. Jest to spowodowane obecnością w tych sygnałach intensywnych pików pochodzących od zanieczyszczeń leku, które nie są obecne w pozostałych próbkach. W celu ułatwienia oceny poprawy jakości sygnałów na Rys. 20 przedstawiono histogramy współczynników korelacji przed i po ich nałożeniu.



Rys. 20 Histogramy współczynników korelacji obliczone między każdym chromatogramem, a sygnałem wzorcowym: (a) przed i (b) po zastosowaniu metody COW

Potencjalne różnice pomiędzy autentycznymi i zafałszowanymi próbkami leku Viagra® badano za pomocą analizy eksploracyjnej z wykorzystaniem metody PCA. Pierwsze trzy czynniki główne opisują ponad 87,86% całkowitej wariancji danych. Projekcja wyników na przestrzeń opisaną przez pierwsze dwa czynniki główne jest przedstawiona na Rys. 21, na którym autentyczne próbki są oznaczone za pomocą symbolu '+', a zafałszowane jako '○'. Analiza projekcji wyników pozwala dostrzec obecność sześciu próbek o odmiennym składzie chemicznym w porównaniu z pozostałymi próbkami. Próbki te pochodzą ze zbioru próbek zafałszowanych i zostały wyłączone ze zbioru modelowego w celu eliminacji ich potencjalnego negatywnego wpływu na konstrukcję reguł dyskryminacyjnych z zastosowaniem metody PLS-DA. Dodatkowo, analiza projekcji wyników uzyskanych dla pierwszych dwóch czynników głównych pozwoliła zaobserwować, że grupa próbek zafałszowanych jest zdecydowanie bardziej niehomogeniczna w porównaniu do grupy próbek autentycznych. Zjawisko to nie jest zaskakujące, gdyż produkcja zafałszowanych leków nie jest objęta wymogami określanymi jako „dobre praktyki produkcji" (z ang. *good manufacturing practises*, GMP), które ściśle precyzują procedury wytwarzania różnego rodzaju produktów w tym także leków. Leki pochodzące z nielegalnych źródeł nie podlegają kontroli jakości czego wynikiem jest ich bogaty skład chemiczny. Duży rozrzut zafałszowanych próbek względem osi PC 2, potwierdza hipotezę, iż nielegalna procedura wytwarzania leków stanowi główne źródło zmienności analizowanych próbek, co może być skutkiem obecności różnego rodzaju zanieczyszczeń w tych lekach. Dodatkowo, projekcja wyników ujawniła tendencję do grupowania się próbek ze względu na ich autentyczność. Największe różnice pomiędzy danymi grupami są spowodowane obecnością zanieczyszczeń, których piki występują przy czasach elucji ok. 2,855 min. i ok. 23,365 min (Rys. 21 c). Natomiast rozrzut próbek względem osi PC 2 powoduje obecność zanieczyszczeń przy czasie elucji ok. 8,8 min. (Rys. 21 d).

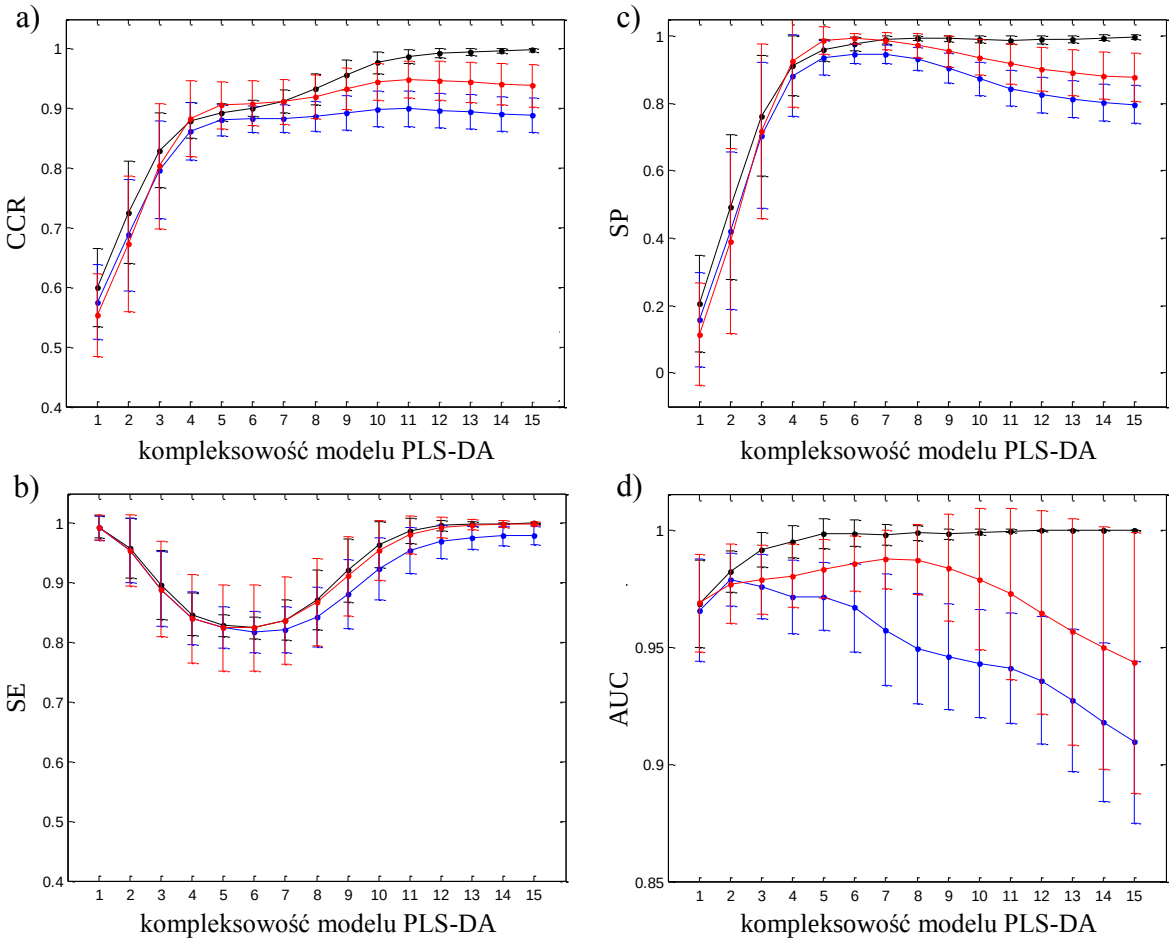


Rys. 21 (a) projekcja wyników uzyskanych dla chromatograficznych profili zanieczyszczeń próbek leku Viagra® na przestrzeń opisaną przez pierwsze dwa czynniki główne (PC 1 - PC 2), 46 próbek autentycznych '+' i 97 zafałszowanych '○', (b) przybliżenie określonego regionu projekcji PC 1 - PC 2, (c) wagi uzyskane dla pierwszego czynnika głównego (PC 1) z wyznaczonymi trzema obszarami czasu elucji (1) 2,855, (2) 8,800 i (3) 23,365 min. oraz (d) wagi uzyskane dla drugiego czynnika głównego (PC 1) z wyznaczonymi czterema obszarami czasu elucji (1) 8,80, (2) 22,02, (3) 23,26, (4) 23,37 min., które odpowiadają pikom substancji różnicującym analizowane próbki

W kolejnym kroku analizy chemometrycznej, na podstawie wstępnie przygotowanych sygnałów chromatograficznych stanowiących profile zanieczyszczeń leku Viagra® skonstruowano wieloparametrowy model dyskryminacyjny. W tym celu wykorzystano dyskryminacyjny wariant metody częściowych najmniejszych kwadratów z walidacją typu Monte Carlo, którą opisano w rozdziale 4.2.

Konstruowany model dyskryminacyjny miał na celu odróżnienie oryginalnych i zafałszowanych próbek leku Viagra®. W pierwszym kroku analizy dyskryminacyjnej ze zbioru wszystkich próbek wydzielono po 35 próbek z każdej grupy. Następnie wybrany

zbiór danych podzielono na zbiór modelowy i wewnętrzny zbiór testowy. Wewnętrzny zbiór modelowy i wewnętrzny zbiór testowy zawierały odpowiednio po 25 i po 10 próbek z każdej z analizowanych grup. Niezależny zbiór testowy, który nie brał udziału w konstrukcji modelu składał się z 8 chromatogramów próbek autentycznych i 8 chromatogramów próbek zafałszowanych. Niezależny zbiór testowy wybrano z wykorzystaniem algorytmu Kennarda i Stona spośród 11 próbek autentycznych i 62 zafałszowanych. Model PLS-DA, skonstruowany dla zbioru modelowego, zwalidowano za pomocą wewnętrznego zbioru testowego i niezależnego zbioru testowego. Cała procedura wyboru próbek do poszczególnych zbiorów, konstrukcja modelu i jego walidacja były powtarzane 1000 razy z zastosowaniem procedury Monte Carlo zgodnie ze schematem przedstawionym na Rys. 15. Taki sposób walidacji pozwolił wykreślić zależności pomiędzy liczbą ukrytych zmiennych, a średnimi wartościami parametrów walidacyjnych modelu takimi jak procent poprawnej klasyfikacji, czułość, specyficzność oraz efektywność przewidywania modelu. Dodatkowo dla każdej wartości parametru wyznaczone zostało odchylenie standardowe. Na podstawie analizy rozkładu uzyskanych błędów dla modeli o różnej kompleksowości wybrano optymalną liczbę ukrytych zmiennych równą 5. Model PLS-DA o optymalnej kompleksowości pozwolił uzyskać procent poprawnej klasyfikacji dla zbioru modelowego na poziomie 89,37% ± 1,48%. Natomiast, dla wewnętrznego zbioru testowego parametr ten wynosi 90,60% ± 3,97%, a dla zewnętrznego zbioru testowego 88,03% ± 2,64%. Ponadto, dla modelu o założonej kompleksowości także pozostałe parametry walidacyjne uzyskały wysokie wartości (Rys. 22). Pozwala to wnioskować, że badany problem dyskryminacyjny może być rozwiązany za pomocą relatywnie prostego liniowego modelu PLS-DA. Średnie wartości wszystkich rozważanych parametrów walidacyjnych modelu wraz z odpowiadającymi im odchyleniami standardowymi zostały przedstawione w Tabeli 4.



Rys. 22 Wykres zależności wartości średnich (a) procentu poprawnej klasyfikacji (CCR), (b) czułości (SE), (c) specyficzności (SP) oraz (d) efektywności modelu (AUC) od kompleksowości modelu PLS-DA z wyznaczonymi wartościami odchyleń standardowych (linie pionowe) określonymi na podstawie procedury Monte Carlo (1000 iteracji) dla wewnętrznego zbioru modelowego (czarna linia), dla wewnętrznego zbioru testowego (czerwona linia) oraz dla zewnętrznego zbioru testowego (niebieska linia)

Pomimo, iż skonstruowany model PLS-DA charakteryzuje się dobrymi wartościami parametrów walidacyjnych, z powodu dużej liczby zmiennych w stosunku do liczby analizowanych próbek zawsze istnieje ryzyko przeuczenia modelu. Rozwiązaniem tego problemu może być ograniczenie liczby zmiennych użytych do modelowania poprzez wybór zmiennych istotnych [37]. W tym celu wykorzystano cztery metody tj. metodę eliminacji zmiennych nieistotnych (UVE) [40], metodę zmiennych znaczących dla projekcji (VIP) [66], współczynnik selektywności (SR) [39] oraz metodę korelacji wieloczynnikowej (SMC) [41]. W połączeniu z metodą PLS-DA pozwoliły one uzyskać modele dyskryminacyjne o zbliżonych wartościach parametrów predykcyjnych. Jednak

w większości przypadków modele skonstruowane na podstawie zmiennych istotnych miały lepsze wartości specyficzności niż modele PLS-DA skonstruowane na podstawie wszystkich zmiennych. Świadczy to o tym, że reguły logiczne skonstruowane dla wybranych zmiennych istotnych lepiej przewidują przynależność zafałszowanych próbek leku Viagra®. Uzyskane wyniki przedstawiono w Tabeli 4. Jak miało to miejsce wcześniej, zastosowano procedurę Monte Carlo zgodnie ze schematem przedstawionym na Rys. 13 stosując liczbę powtórzeń równą 1000.

W metodzie UVE-PLS-DA do każdego z wylosowanych zbiorów modelowych dodana została macierz zmiennych nieistotnych (10 000 zmiennych dla każdego sygnału). Zmienne nieistotne stanowiły liczby losowe wybrane z rozkładu normalnego i pomnożone przez współczynnik $c = 10^{-12}$. Następnie, zbudowano model PLS-DA, a jako istotne zmienne wybierano te których wartość bezwzględna stabilności liczona na podstawie uzyskanych współczynników regresji była większa niż 99% maksymalnej wartości stabilności uzyskanej dla zmiennych nieistotnych. Zmienne istotne były wybierane w każdym kroku procedury Monte Carlo, co w rezultacie dało 1000 zbiorów zmiennych istotnych. Końcowy zestaw wybranych zmiennych istotnych zawierał tylko te, które były uznawane za istotne w każdej z iteracji (zmienne, których częstotliwość wyboru wynosiła 100%).

W metodzie VIP założona wartość graniczna parametru określającego istotność zmiennych była równa 1, a końcowy model zawierał zmienne uznawane za istotne w każdym z powtórzeń procedury Monte Carlo. Cały proces wyboru zmiennych za pomocą metody VIP był powtórzony 3 razy, przy czym, w każdej kolejnej iteracji wykorzystano jedynie te zmienne, które były uznawane za istotne w poprzedniej procedurze VIP ze 100% częstotliwością wyboru.

W metodzie SR granica istotności odpowiadała parametrowi SR równemu 0,9, a końcowy model zawierał zmienne, które były określane jako istotne w 95% iteracji Monte Carlo.

Podobnie stosując metodę SMC do końcowego zbioru modelowego wybrano tylko te zmienne, które były określane jako istotne w każdej iteracji procedury Monte Carlo.

Analiza wyników przedstawionych w Tabeli 4 pozwala stwierdzić, iż każda z metod wyboru zmiennych prowadzi do obniżenia kompleksowości, a tym samym do obniżenia ryzyka przeuczenia modelu PLS-DA. Modele dyskryminacyjne skonstruowane dla danych zawierających zmienne wybrane za pomocą metody SR i SMC miały kompleksowość niższą o jeden w porównaniu z modelem otrzymanym dla wszystkich zmiennych. Pozostałe metody

UVE i VIP zredukowały kompleksowość modeli skonstruowanych z ich zastosowaniem do 2 czynników. Należy podkreślić, że zmniejszenie kompleksowości modeli konstruowanych dla analizowanych danych nie spowodowało pogorszenia się ich parametrów walidacyjnych. Największą kompresję liczby zmiennych uzyskano za pomocą metody SR. Z 13 291 oryginalnych zmiennych jedynie 21 zostało uznanych jako istotne do konstrukcji modelu pozwalającego rozróżnić zafałszowane i autentyczne próbki leku Viagra®. Najwięcej zmiennych istotnych zidentyfikowano z użyciem metody SMC. Z 13 291 zostało wybranych 3 641 zmiennych.

Różna liczba zmiennych wybranych przez zastosowane podejścia wynika z odmiennych kryteriów jakie są stosowane w wykorzystanych metodach wyboru zmiennych (3.2.2). Podczas, gdy współczynnik selektywności, SR definiuje istotność zmiennej na podstawie wariancji danych jaka jest przez nią opisywana, metoda zmiennych znaczących dla projekcji, VIP określa, które zmienne są istotne bazując na tym jak dobrze dana zmienna opisuje nie tylko wariancje danych ale także kowariancje pomiędzy zbiorem danych a zmienna zależną.

Wszystkie skonstruowane modele PLS-DA bardzo dobrze przewidują przynależność próbek do odpowiednich grup dla zewnętrznego zbioru testowego o czym świadczą średnie wartości procentu poprawnej klasyfikacji, które wynoszą powyżej 88%. Dodatkowo, wartości odchyleń standardowych dla danych wartości współczynnika CCR są poniżej 2,86%.

Model skonstruowany z wykorzystaniem metody VIP-PLS-DA pozwala bardzo dobrze przewidzieć przynależność próbek zewnętrznego zbioru testowego, jednak charakteryzuje się on lepszym przewidywaniem próbek autentycznych niż zafałszowanych dla zewnętrznego zbioru testowego o czym świadczą uzyskane wyższe wartości parametru opisującego poprawność przewidywania próbek zafałszowanych (specyficzności) niż wartości parametru poprawności przewidywania próbek autentycznych (czułości) (SE = 98,69% ± 1,38%; SP = 94,16% ± 3,52%). Pozostałe modele dyskryminacyjne wykazywały odwrotną tendencję, a mianowicie lepiej przewidywały przynależność próbek autentycznych niż zafałszowanych dla zewnętrznego zbioru testowego (zob. Tabela 4).

Przedstawione wyniki badań potwierdzają użyteczność proponowanej we wcześniejszym rozdziale metody walidacji połączonej z analizą PLS-DA oraz różnymi metodami wyboru zmiennych do weryfikacji autentyczności leku Viagra®. Wyznaczenie zmiennych istotnych do budowy modelu nie tylko zniwelowało ryzyko przeuczenia modelu, ale także pozwoliło zidentyfikować obszary czasu elucji, którym odpowiadają piki pochodzące od substancji

stanowiących potencjalne markery nielegalnego procederu fałszowania leków. Jednak identyfikacja tych substancji wymaga zastosowania komplementarnych technik analitycznych takich jak np. HPLC-MS.

Więcej szczegółów dotyczących identyfikacji zafałszowań leku Viagra® znajduje się w publikacji „The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles", Analyst, 141 (2016) 1060-1070, która stanowi Załącznik nr 2 do niniejszej rozprawy doktorskiej.

Tabela 4 Wyniki uzyskane dla modeli dyskryminacyjnych PLS-DA bez i z zastosowaniem metod wyboru zmiennych (wartości wyrażone w %) (UVE – metoda eliminacji zmiennych nieistotnych, VIP – metoda zmiennych znaczących dla projekcji, SR – współczynnik selektywności, SMC – metoda korelacji wieloczynnikowej, n – liczba zmiennych, f – liczba czynników PLS-DA, CCR – procent poprawnej klasyfikacji, SE – czułość, SP – specyficzność, parametr AUC)

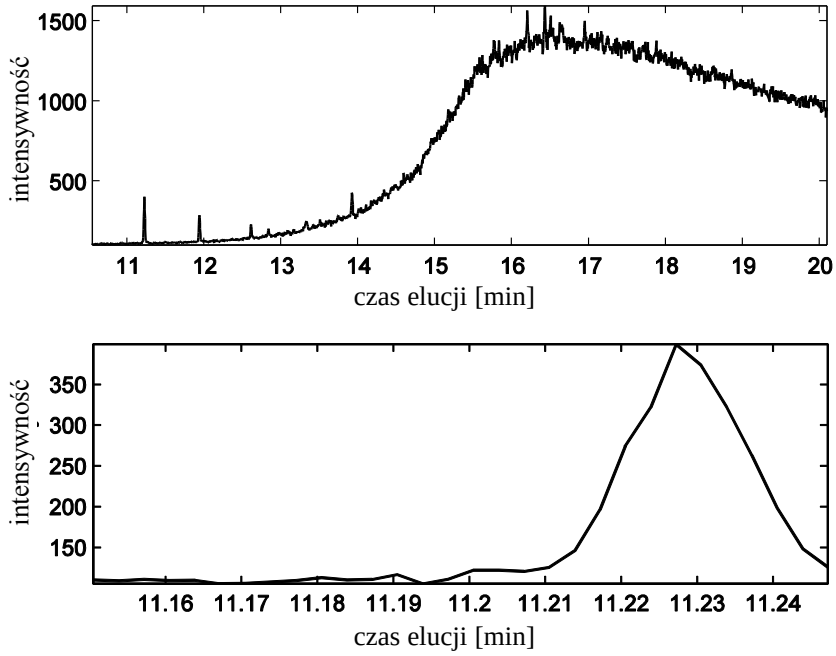
Model	f	n	Zbiór modelowy (Monte Carlo)				Zbiór testowy (Monte Carlo)				Niezależny zbiór testowy			
			CCR	SE	SP	AUC	CCR	SE	SP	AUC	CCR	SE	SP	AUC
PLS-DA	5	13 291	89,37 ± 1,48	82,82 ± 1,88	95,92 ± 3,25	99,90 ± 0,60	90,60 ± 3,97	82,44 ± 7,31	98,75 ± 4,06	98,40 ± 1,30	88,03 ± 2,64	82,48 ± 3,48	93,58 ± 5,03	97,20 ± 1,50
UVE	2	674	89,11 ± 1,44	82,47 ± 1,44	95,76 ± 2,62	96,00 ± 1,30	90,74 ± 3,70	81,64 ± 7,39	99,83 ± 0,56	89,10 ± 5,40	88,36 ± 2,19	82,45 ± 3,28	94,28 ± 3,40	94,00 ± 2,60
VIP	2	83	97,84 ± 1,39	99,10 ± 1,04	99,57 ± 2,26	99,90 ± 0,01	93,34 ± 3,32	100,00 ± 0,00	86,68 ± 6,65	95,60 ± 4,70	96,42 ± 2,04	98,69 ± 1,38	94,16 ± 3,52	98,20 ± 1,70
SR	4	21	91,32 ± 0,69	82,65 ± 1,37	100,00 ± 0,00	95,50 ± 0,80	90,71 ± 3,73	81,64 ± 7,39	99,79 ± 0,65	88,90 ± 4,60	89,72 ± 1,90	81,26 ± 3,40	98,19 ± 1,60	93,60 ± 2,20
SMC	4	3 641	94,22 ± 1,95	91,47 ± 3,76	96,97 ± 2,79	98,60 ± 0,90	94,41 ± 3,11	90,30 ± 6,39	98,52 ± 2,58	99,90 ± 0,30	91,38 ± 2,86	88,71 ± 5,44	94,05 ± 4,35	96,20 ± 1,60

4.4 Identyfikacja skażenia wody tributylocyną

Analiza próbek środowiskowych, ze względu na ich złożony skład jest zazwyczaj kosztowna i czasochłonna. Dlatego stale poszukuje się nowych procedur analitycznych ułatwiających analitykowi identyfikację substancji zawartych w próbkach o złożonej matrycy. Przykładem rutynowo wykonywanej analizy środowiskowej jest kontrola jakości wód rzek i jezior. Jedną z substancji, która wymaga stałej oceny jej obecności w wodach lądowych jest tributylocyna (z ang. *tributyltin*, TBT). Jest to środek biobójczy, szeroko stosowany jako składnik farb przeciwporostowych, w których obecność TBT powodowała zapobieganie lub spowolnienie wzrostu organizmów na zabezpieczonych powierzchniach. Tego typu środki stosowano głównie w przemyśle stoczniowym. Początkowo farby przeciwporostowe zawierające TBT były uważane za ekologiczne i bezpieczne dla środowiska, jednak z czasem dowiedziono, że tributylocyna uwalnia się do wód, powodując ich toksyczne skażenie. Z tego powodu wprowadzono międzynarodowe przepisy zabraniające stosowania produktów zawierających TBT, mające za zadanie ograniczenie postępującego skażenia wód przez ten związek oraz produkty jego degradacji [67,68]. Ze względu na stosunkowo długi okres półtrwania TBT, który zależy od takich czynników jak źródło pochodzenia i warunki środowiskowe, toksyczne efekty działania tej substancji są nadal zauważalne [69,70]. Z tego powodu obecność TBT w próbkach środowiskowych, a w szczególności w próbkach wód morskich i lądowych wymaga stałej kontroli. Wody lądowe ze względu na złożony skład są z reguły analizowane za pomocą technik chromatograficznych. Sygnały chromatograficzne uzyskane dla próbek środowiskowych charakteryzują się dużą liczbą pików, które mogą się nakładać się na siebie co komplikuje pozyskiwanie informacji analitycznej. Dodatkowo na jakość chromatogramów wpływa występowanie przesunięć pików. W niniejszych badaniach zaproponowano modele diagnostyczne pozwalające ocenić obecność tributylocyny w wodzie bez konieczności wykonywania oceny ilościowej. Modele diagnostyczne stanowiące część opracowanego systemu eksperckiego były konstruowane z wykorzystaniem narzędzi chemometrycznych w oparciu o sygnały chromatograficzne stanowiące chemiczne odciski palca badanych próbek. Podejście to pozwoliło na uzyskanie maksimum informacji na temat składu badanych próbek. Analizie poddano próbki wód lądowych, które zostały pobrane na zlecenie Głównego Inspektoratu Ochrony Środowiska w związku z prowadzonymi na szeroką skalę badaniami jakości wód lądowych. Jeden z etapów tych badań obejmował wykrywanie TBT w wodzie. Całość eksperymentu została przeprowadzona w akredytowanym laboratorium firmy Polcargo International w Szczecinie. Zgodnie z normą

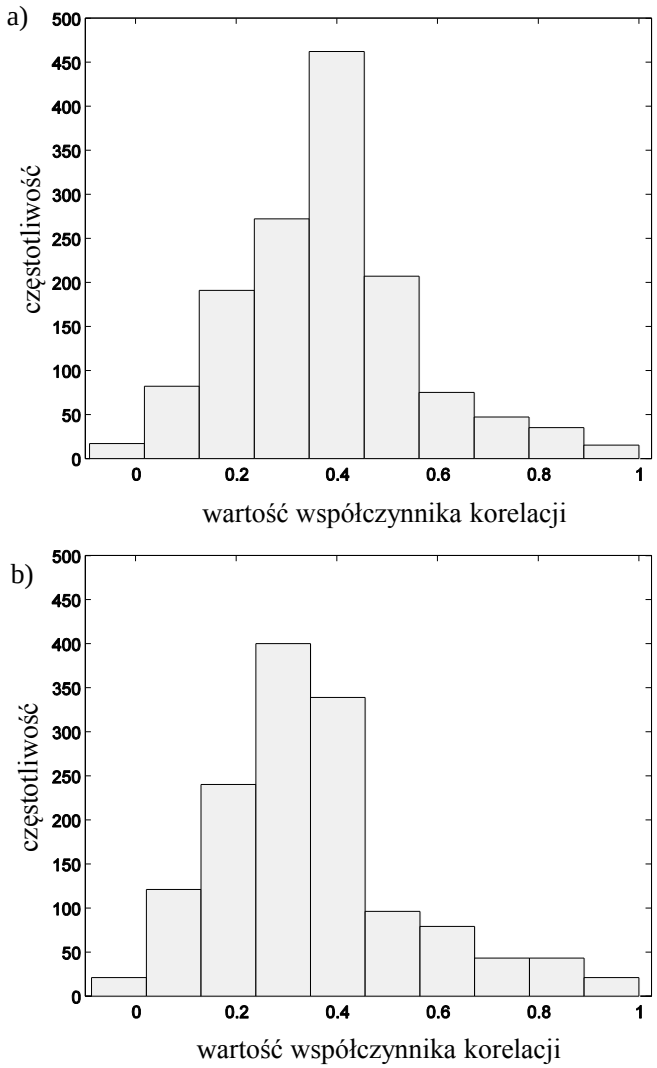
PN-EN ISO 17353: 2006 przeanalizowano 1403 próbek wody pobranych w latach 2011 i 2013. Do ustalenia obecności TBT wykorzystano chromatografię gazową sprzężoną ze spektrometrem mas, GC-MS. Uzyskane sygnały stanowiące chromatograficzne odciski palca próbek wody, pomimo rygorystycznych wymogów dotyczących warunków prowadzenia analizy chromatograficznej, zawierały takie składniki jak szum, linia podstawowa oraz przesunięcia pików pochodzących od tych samych substancji. Komponenty te mogą fałszować wyniki analizy chemometrycznej, a tym samym negatywnie wpływać na proces wnioskowania. Kształt linii podstawowej różnił się dla poszczególnych próbek, natomiast dla pojedynczych sygnałów zauważono prawidłowość polegającą na wzroście intensywności linii podstawowej wraz z czasem prowadzenia analizy chromatograficznej. Obszary sygnału o wysokiej intensywności pików charakteryzowały się wyższym poziomem szumu, niż obszary o niskiej intensywności sygnału. Dodatkowo, w analizowanych chromatograficznych odciskach palca zauważono przesunięcia pików względem siebie. Sygnały GC-MS zostały wstępnie przygotowane poprzez eliminację linii podstawowej, usunięcie szumu oraz korekcję przesunięć pomiędzy pikami. Eliminacji szumu dokonano za pomocą prostej transformacji logarytmicznej (log10) [7,9]. Linia podstawowa została usunięta za pomocą metody częściowych najmniejszych kwadratów z funkcja kary, PAsLS. Najlepsze rezultaty uzyskano dla parametrów wejściowych $p = 10^{-4}$ i $\lambda = 10^{4}$. Następnie w analizowanych sygnałach usunięto przesunięcia odpowiadających sobie pików za pomocą metody COW, w której sygnał wzorcowy wybrano zgodnie z założeniami opisanymi w [14,15]. Jako sygnał wzorcowy względem, którego były eliminowane przesunięcia pików został wybrany sygnał posiadający najlepszą korelację względem pozostałych sygnałów chromatograficznych. Wybrany sygnał pochodził z grupy próbek, które nie w zawierały TBT. Lepsze wyniki eliminacji przesunięć pików uzyskano dla pików znajdujących się poza obszarem obejmującym pik pochodzący od kationu tributylocyny. Jest to prawdopodobnie spowodowane brakiem danego piku w sygnale wzorcowym. Aby zbadać ewentualne zwiększenie wydajności wyrównania sygnałów, zastosowano to samo podejście wybierając jako sygnał wzorcowy chromatogram z grupy próbek zawierających TBT. W związku z czym na etapie eliminacji przesunięć pomiędzy pikami przebadano dwa sygnały wzorcowe pochodzące z dwóch różnych grup oraz różne ustawienia parametrów wejściowych (długość sekcji i parametr elastyczności). Ze wszystkich ustawień parametrów wejściowych szczegółowo przeanalizowano dwie pary o skrajnych wartościach – krótka sekcja (20 punktów pomiarowych) i dłuższa sekcja (28 punktów pomiarowych), jak również mała wartość parametru elastyczności (2) oraz większa wartość parametru elastyczności (4).

Przykładowy surowy chromatograficzny odcisk palca otrzymany dla analizowanej próbki wody przedstawiono na Rys. 23.



Rys. 23 Przykładowy chromatograficzny odcisk palca otrzymany dla próbki wody zawierającej TBT z powiększonym regionem odpowiadającym czasowi elucji badanego analitu

Histogramy uzyskanych współczynniki korelacji sygnału wzorcowego z pozostałymi sygnałami chromatograficznymi przedstawiono na Rys. 24. Analizowane sygnały chromatograficzne różnią się znacznie pomiędzy sobą w związku z czym ich podobieństwo względem poszczególnych sygnałów wzorcowych jest stosunkowo małe. Histogram przedstawiony na Rys. 24 a wskazuje, że dla ponad 450 chromatograficznych odcisków palca uzyskano średnie wartości współczynników korelacji o wartości powyżej 0,4 w odniesieniu do sygnału wzorcowego. Mniejsza liczba chromatogramów wykazuje wartości współczynników korelacji wyższe od 0,4, w przypadku gdy przesunięcia pomiędzy pikami były eliminowane względem chromatogramu z grupy sygnałów uzyskanych dla próbek zawierających TBT (patrz Rys. 24 b).



Rys. 24 Histogramy współczynników korelacji, które zostały obliczone pomiędzy sygnałami chromatograficznymi, a sygnałem odniesienia wybranym z grupy próbek:
(a) niezawierających TBT oraz (b) zawierających TBT

Aby porównać wyniki działania metody COW dla różnych parametrów wejściowych (długość sekcji i parametr elastyczności) wyznaczono wartości współczynników korelacji pomiędzy sygnałem wzorcowym i pozostałymi chromatogramami przed i po eliminacji przesunięć. Suma różnic współczynników korelacji została oznaczona jako AG, co pozwoliło w prosty sposób scharakteryzować wydajność z jaką usunięto przesunięcia pików chromatograficznych. Zauważono, że niezależnie od wybranych wartości parametrów wejściowych i wybranego sygnału wzorcowego wartości współczynników korelacji po

nałożeniu sygnałów za pomocą metody COW wzrosły. Najlepsze wyniki eliminacji przesunięć pików (AG = 114,5) uzyskano dla sekcji zawierającej dwadzieścia punktów pomiarowych (N = 20) i parametru elastyczności równego cztery (t = 4). Wartości parametrów charakteryzujących efektywność nakładania sygnałów za pomocą metody COW przedstawiono w Tabeli 5.

Tabela 5. Wyniki eliminacji przesunięć pomiędzy pikami z zastosowaniem metody COW dla przedstawionych parametrów wejściowych (długości sekcji (N) i parametru elastyczności (t)) przedstawione jako różnice współczynników korelacji sygnałów przed i po nałożeniu sygnałów (AG)

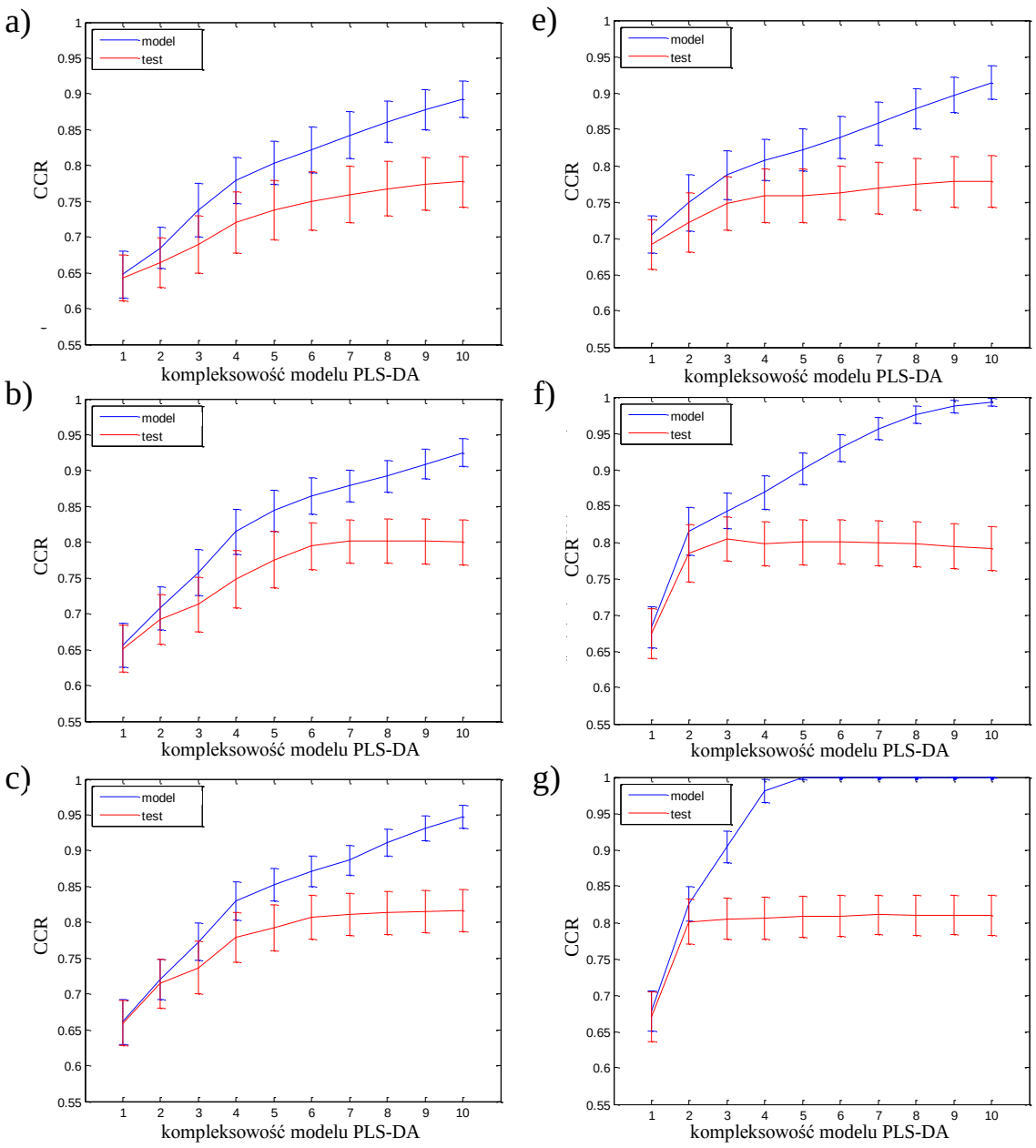
Sygnał wzorcowy	Lp.	N	t	ΣΔ(+)	ΣΔ(-)	AG
Wybrany z grupy nie zawierającej TBT	1	28	2	+ 62,5717	- 2,8840	+ 69,6877
	2	28	4	+ 78,2420	- 3,0301	+ 75,2119
	3	20	2	+ 73,0242	- 2,5966	+ 70,4276
	4	20	4	+ 107,5005	- 4,5093	+ 102,9912
Wybrany z grupy zawierającej TBT	1	28	2	+ 80,0706	- 2,5456	+ 77,5250
	2	28	4	+ 103,6455	- 3,5887	+ 100,0568
	3	20	2	+ 89,8385	- 2,9274	+ 86,9111
	4	20	4	+ 119,2912	- 4,7873	+ 114,5039

Wstępnie przygotowane chromatograficzne odciski palca, odpowiadające zakresowi czasu elucji od 10,55 min. do 20,10 min., wykorzystano do konstruowania modeli dyskryminacyjnych służących do rozróżnienia analizowanych grup próbek ze względu na obecność w nich TBT. W celu opracowania reguł dyskryminacyjnych wykorzystano dyskryminacyjny wariant metody częściowych najmniejszych kwadratów PLS-DA [29,71]. Poprawność przewidywania skonstruowanych modeli sprawdzono poprzez określenie ich

parametrów walidacyjnych takich jak czułość, specyficzność, efektywność oraz procent poprawnej klasyfikacji. Dodatkowo, w celu uzyskania jak najbardziej wiarygodnych wyników wykorzystano do konstrukcji modeli dyskryminacyjnych podejście typu Monte Carlo [31]. Pozwoliło to estymować zmienność danych oraz wyznaczyć niepewności pomiaru dla uzyskanych wartości parametrów. Modele dyskryminacyjne były konstruowane dla oryginalnych danych oraz dla danych, które wstępnie przygotowano wykorzystując różnego rodzaju podejścia chemometryczne. Pozwoliło to określić wpływ wykorzystanych metod wstępnego przygotowania danych na wyniki dyskryminacji.

Analizowane metody przygotowania danych obejmowały wykorzystanie dwóch metod mających za zadanie poprawę stosunku sygnału do szumu. Pierwsza z nich polegała na zastosowaniu pierwiastka z kwadratu sygnału, natomiast druga to transformacja logarytmiczna (log10). Dodatkowo, przebadano wpływ eliminacji linii podstawowej na wyniki uzyskane za pomocą modelu PLS-DA. Aby możliwe było porównanie uzyskanych wyników wszystkie modele zostały zbudowane przy użyciu tego samego schematu. Analizowany zbiór danych zawierał dwie grupy próbek: z TBT (157 próbek) oraz bez TBT (1 246 próbek). W celu symulacji zmienności zbioru modelowego zastosowano podejście Monte Carlo, z liczbą iteracji równą 500. Wszystkie zbiory zarówno modelowe jak i testowe były zbalansowane. Zbiór modelowy zawsze był konstruowany z losowo wybranych 158 próbek (po 79 próbek z analizowanych grup). Natomiast zbiór testowy zawierał 156 próbek (po 78 próbek z każdej grupy). Próbki do zbioru testowego były wybierane losowo ze zbioru danych po wyłączeniu zbioru modelowego. W ten sposób pojedynczy zestaw danych składał się z dwóch wzajemnie wykluczających się zbiorów danych. Na podstawie każdego zbioru modelowego konstruowano model PLS-DA o kompleksowości f = 1, 2, ..., 10. Poprawność przewidywania modeli została przedstawiona za pomocą wykresu zależność średniej wartości współczynnika poprawności przewidywania modelu od jego kompleksowości. Zależność ta była wyznaczana zarówno dla wszystkich zbiorów modelowych jak i dla zbiorów testowych. Dla wszystkich wartości średnich współczynnika poprawności przewidywania modelu wyznaczono odchylenie standardowe na podstawie uzyskanych 500 wartości danego parametru. Wyniki analizy dyskryminacyjnej danych przed i po wstępnym przygotowaniu ilustruje Rys. 25. Odchylenie standardowe uzyskanych wartości współczynnika CCR jest przedstawione jako pionowa linia wyznaczająca możliwy zakres błędu danej wartości parametru. Uważna analiza uzyskanych wyników wskazała kilka ciekawych zależności. W przypadku modeli konstruowanych na podstawie surowych danych

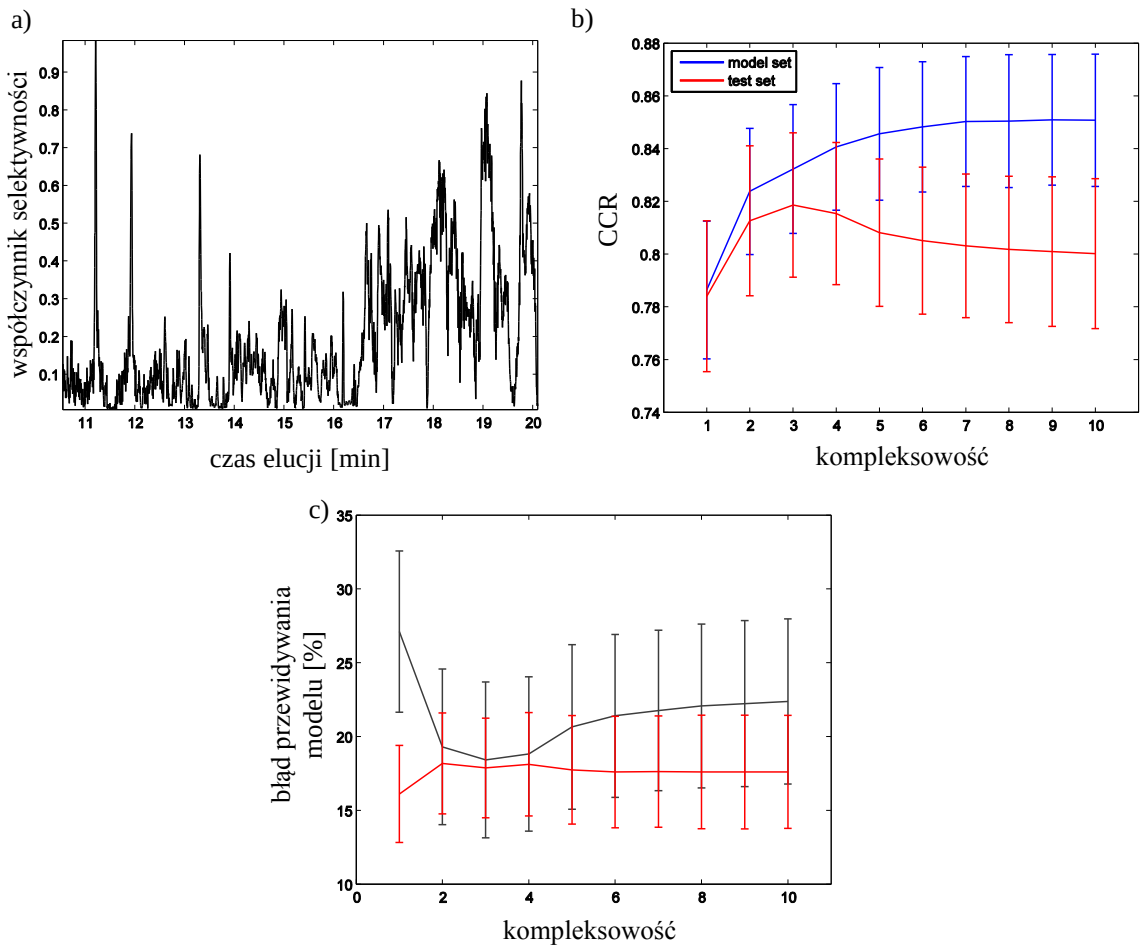
chromatograficznych o małej kompleksowości otrzymano, względnie słabe wyniki, które polepszały się wraz ze zwiększaniem liczby czynników PLS-DA. Dla prostych modeli (f = 3) uzyskano współczynniki poprawności przewidywania poniżej 0,75 zarówno dla zbioru modelowego jak i testowego. Z drugiej strony, stosunkowo duża złożoność modelu może być traktowana jako próba zrekompensowania błędnych i nieprecyzyjnych informacji związanych z obecnością TBT w analizowanych danych. Wykonane przekształcenia związane ze wstępnym przygotowaniem danych spowodowały nieznaczną poprawę zdolności predykcyjnych modelu. Najlepsze wyniki uzyskano bazując na chromatograficznych odciskach palca poddanych transformacji logarytmicznej, dla których skonstruowany model uzyskał współczynnik poprawności przewidywania ponad 0,7 zarówno dla zbioru testowego jak i dla zbioru modelowego dla dwóch czynników PLS-DA. Kolejna procedura polegająca na usunięciu linii podstawowej z analizowanych sygnałów chromatograficznych znacząco poprawiła zdolności predykcyjne konstruowanych modeli dyskryminacyjnych. Również w tym przypadku procent poprawnej dyskryminacji zwiększał się systematycznie ze wzrostem kompleksowości modelu. Modele skonstruowane dla danych po wstępnym przygotowaniu wykazują procent poprawnej dyskryminacji dla zbioru testowego na poziomie 80% dla trzech czynników PLS niezależnie od metody jaka była zastosowana do eliminacji szumu.



Rys. 25 (a) wyniki modeli PLS-DA skonstruowanych dla surowych sygnałów chromatograficznych oraz sygnałów wstępnie przygotowanych za mocą różnych metod takich jak, (b) normalizacja, (c) transformacja logarytmiczna (log10), (d) eliminacja linii podstawowej za pomocą metody PAsLS, (e) eliminacja linii podstawowej za pomocą metody PAsLS i normalizacja oraz (f) eliminacja linii podstawowej za pomocą metody PAsLS i transformacja logarytmiczna (log10)

W kolejnym etapie badań obejmującym opracowanie systemu eksperckiego pozwalającego usprawnić kontrolę jakości próbek wody pod względem obecności w nich tributylocyny zaproponowano wykorzystanie współczynnika selektywności, SR. Określenie istotnych

zmiennych w kontekście konstrukcji modeli diagnostycznych, pozwala na eliminację ryzyka przeuczenia modelu w związku z dużą liczbą zmiennych w stosunku do liczby analizowanych próbek. Metoda współczynnika selektywności jest szczegółowo opisana w podrozdziale 3.2.2 oraz w [38,39]. W przypadku analizy sygnałów chromatograficznych wybór zmiennych istotnych pozwala zidentyfikować odpowiednie czasy elucji przy których wymywane są związki odpowiedzialne za różnice pomiędzy analizowanymi grupami próbek. Do wyznaczenia współczynników selektywności analizowanych zmiennych na podstawie skonstruowanych modeli PLS-DA zastosowano procedurę Monte Carlo. Dzięki temu możliwa była estymacja współczynnika SR dla każdej zmiennej oraz określenie odchyleń standardowych dla uzyskanych wartości (zob. Rys. 26). Można zaobserwować, że pomiędzy 11 i 12 min. znajdują się zmienne których wartości współczynnika selektywności są większe od 1, które z reguły są uznawane za istotne do konstrukcji modelu. Należy zaznaczyć, że każda zmienna tak naprawdę odpowiada frakcji eluatu wymywanej w określonym czasie. Oznacza to, że w wyznaczonym przedziale czasu elucji wymywane są substancje odpowiedzialne za różnicowanie badanych dwóch grup próbek wody (zawierających i niezawierających TBT). Poprawność tej hipotezy potwierdza fakt, że w rzeczywistości pik chromatograficzny pochodzący od kationu tributylocyny występuje w danym przedziale czasu elucji. Model PLS-DA skonstruowany na podstawie wybranych 20 istotnych zmiennych z wykorzystaniem trzech czynników PLS posiadał wartość współczynnika CCR ok. 0,82 (patrz Rys. 26 b). Tym samym można stwierdzić, że bazując na dużo mniejszej liczbie zmiennych w porównaniu do wyjściowego zestawu danych uzyskano podobne wyniki poprawności przewidywania dla modelu dyskryminacyjnego. Dodatkowo wyznaczono procent próbek wody nieprawidłowo sklasyfikowanych przez dany model dyskryminacyjny, który był określany dla każdej z grup oddzielnie (Rys. 26 c). Interesujący jest fakt, że dla każdej z grup uzyskano zarówno podobne poziomy błędu około 0,19 jak i porównywalne zakresy niepewności.



Rys. 26 (a) wartości współczynnika selektywności dla analizowanych zmiennych, (b) właściwości predykcyjne modelu PLS-DA który został skonstruowany na podstawie wybranych zmiennych istotnych (20 zmiennych o współczynniku selektywności powyżej jedynki) wyrażone jako procent poprawnej dyskryminacji wyznaczony dla zbioru modelowego i testowego w funkcji kompleksowości modelu (niepewności wyznaczono metodą Mont Carlo) i (c) błąd przewidywania modelu dla różnych wartości kompleksowości wyznaczony oddzielnie dla każdej grupy próbek wody (odchylenia standardowe wyznaczono z wykorzystaniem metody typu Monte Carlo)

Pozostałe parametry walidacyjne uzyskane dla modelu PLS-DA skonstruowanego dla całych sygnałów chromatograficznych oraz dla modelu zbudowanego na podstawie wybranych zmiennych również miały podobne wartości. W tabeli poniżej przedstawiono uzyskane wyniki procentu poprawnej klasyfikacji (CCR), czułości (SE) i specyficzności (SP) dla modeli skonstruowanych na podstawie całych sygnałów instrumentalnych oraz dla zbioru danych zawierającego jedynie istotne zmienne z jednoczesnym uwzględnieniem zastosowania metod wstępnego przygotowania danych.

Tabela 6. Wyniki uzyskane dla modeli PLS-DA wyrażone jako procent poprawnej klasyfikacji (CCR), czułość i specyficzność uzyskane dla niezależnych zbiorów testowych. Modele uzyskano dla surowych danych i danych po wstępnym przygotowaniu obejmującym korektę linii podstawowej, transformację log10 i nakładanie sygnałów metodą COW. Niepewności pomiarów dla wyznaczonych wartości średnich parametrów walidacyjnych modeli otrzymano stosując procedurę Monte Carlo (500 powtórzeń). Gwiazdką (*) oznaczone są wyniki uzyskane dla modeli PLS-DA skonstruowanych dla zmiennych istotnych wyznaczonych za pomocą współczynnika selektywności (SR).

Typ danych	Model	CCR	Czułość	Specyficzność
Surowe dane	PLS-DA	0,805 ± 0,028	0,819 ± 0,046	0,791 ± 0,059
	PLS-DA*	0,813 ± 0,028	0,818 ± 0,034	0,807 ± 0,053
Dane po wstępnym przygotowaniu	PLS-DA	0,795 ± 0,030	0,800 ± 0,048	0,790 ± 0,054
	PLS-DA*	0,793 ± 0,029	0,841 ± 0,038	0,655 ± 0,057

Przeprowadzone badania pozwoliły zweryfikować możliwość oceny próbek wody lądowej pod względem obecności tributylocyny na podstawie chromatograficznych odcisków palca. Wszystkie badane chromatogramy były rejestrowane w tym samym zakresie czasu elucji od 10,55 min. do 20,10 min. Dlatego też można wnioskować, że dyskryminacja tego rodzaju próbek bazuje na istotnych regionach chromatograficznych odcisków palca, które ilustrują największe różnice w składzie wody pomiędzy analizowanymi grupami. Ponadto, obecność lub brak TBT w próbkach potencjalnie może korelować z obecnością lub brakiem innych substancji chemicznych. W związku z tym, wykorzystanie chemicznych odcisków palca może ułatwić dyskryminację próbek. Zaimplementowanie prostej metody dyskryminacyjnej do badania obecności TBT na podstawie sygnałów chromatograficznych stanowi jeden z etapów konstrukcji wiarygodnego systemu eksperckiego opracowanego przy użyciu metody uczenia maszynowego. Proponowane podejście pozwala uzyskać relatywnie dobrą dyskryminację próbek dla surowych danych chromatograficznych, które zostały uzyskane bezpośrednio w toku rutynowego monitorowania zanieczyszczenia środowiska tributylocyną. Stanowi to obiecującą prognozę zastosowania proponowanej metodyki do usprawnienia

badań dotyczących obecności TBC w próbkach wód lądowych. Natomiast fakt, że do badań wykorzystano chromatograficzne odciski palca badanych próbek pozwala wnioskować, że zawierają one istotne informacje dotyczące obecności TBT oraz opisują różne źródła zmienności danych, które mogą wpływać na dyskryminację próbek. Na zdolność przewidywania konstruowanych modeli wpływa ogólna jakość sygnałów instrumentalnych. Zbudowany system ekspercki może być wykorzystywany jako dodatkowe wsparcie dla podejmowania decyzji przez personel laboratoryjny, co pozwoli skrócić czas prowadzenia analizy. Stosunkowo wysoka wydajność przewidywania skonstruowanych modeli dyskryminacyjnych (procent poprawnej klasyfikacji ok. 0,8, a czułość i specyficzność ok. 0,8) wykazuje, że chromatograficzne odciski palca uzyskane dla próbek wody zawierających i nie zawierających tributylocynę stanowią bazę próbek dla dalszego rozwoju badań. Modelowanie tejże bazy sygnałów instrumentalnych otwiera możliwości budowy systemu eksperckiego opierającego się na zasadach logiki, skonstruowanych w oparciu o inne metody uczenia maszynowego co zostało omówione w [72]. Za pomocą modeli dyskryminacyjnych skonstruowanych na podstawie analizowanych danych istnieje możliwość potwierdzenia obecności TBT w nowych próbkach z co najmniej 80% prawdopodobieństwem. Jednak należy pamiętać, że dane modele dyskryminacyjne mogą być wykorzystywane dla sygnałów instrumentalnych uzyskanych dla próbek polskich wód śródlądowych, które są analizowane zgodnie z polską normą PN-EN ISO 17353:2006. Wyniki te dowodzą, że uzyskana baza danych zawiera istotne informacje i ukazuje stosunkowo dużą różnorodność próbek polskich wód śródlądowych. Oczywiście, opracowanie danych metod statystycznych nie pozwala na to, aby personel laboratoryjny pominął procedurę kalibracji podczas rutynowej analizy próbek wody. Dlatego należy pamiętać, że podejmowanie decyzji w kontekście prowadzonych badań wymaga analizy sygnałów instrumentalnych, których jakość w dużym stopniu zależy od doświadczenia i dokładności wykonywania analiz przez pracowników laboratorium. Przedstawiona metodyka wykorzystania narzędzi chemometrycznych do opracowania systemu eksperckiego może być również wykorzystana w kontekście rutynowego monitorowania innych substancji priorytetowych w różnego rodzaju próbkach.

Więcej szczegółów dotyczących identyfikacji skażenia wody tributylocyną znajduje się w publikacji „Expert system for monitoring the tributyltin content in inland water samples”, Chemometrics and Intelligent Laboratory Systems, 149 (2015) 123-131, która stanowi Załącznik nr 3 do niniejszej rozprawy doktorskiej.

4.5 Metody badania autentyczności leków

W ostatnich latach zaobserwowano znaczny wzrost liczby przypadków zafałszowań leków. Może być to spowodowane łatwym dostępem do nowoczesnych technologii, które mogą być wykorzystywane do „kopiowania" leków oraz brakiem odpowiedniej kontroli nad produktami farmaceutycznymi, które są sprzedawane przez internet [73]. Niemożliwe jest uzyskanie dokładnych danych na temat skali fałszowania leków. Przyjmuje się, że 10% leków na rynku światowym to leki podrobione. Oczywiście dany udział procentowy jest różny dla poszczególnych krajów. W krajach wysokorozwiniętych, fałszowane leki stanowią około 1% całkowitej liczby kontrolowanych leków. Ponad 50% leków sprzedawanych przez internet, jest zafałszowana lub są to leki o niskiej jakości. Do najczęściej fałszowanych leków należą środki przeciwdrobnoustrojowe (28%), hormony (22%), leki przeciwhistaminowe (17%), środki rozszerzające naczynia krwionośne (7%), leki na zaburzenia erekcji (5%) oraz leki przeciwdrgawkowe (2%) [74]. Ze względu na poważne zagrożenie związane z zażywaniem zafałszowanych leków oraz możliwość szybkiego transportu tego typu farmaceutyków pomiędzy krajami, konieczne jest opracowanie nowych, stosunkowo prostych i efektywnych metod, które będą wspierać proces kontroli jakości leków. Leki autentyczne zwykle mogą być odróżnione od zafałszowanych przez analizę ich składu chemicznego [75]. Substancje zawarte w lekach mające właściwości lecznicze stanowią tak zwane składniki aktywne leków (API). Ze względu na stężenie API leki można podzielić na cztery grupy:

- leki zawierające poprawny związek stanowiący API, którego zawartość jest zgodna z deklaracją producenta,
- leki zawierające poprawny związek stanowiący API, jednak jego dawka jest nieprawidłowa,
- leki zawierające niepoprawną substancję stanowiącą API,
- leki nie posiadające API (placebo).

W wielu przypadkach, identyfikacja autentyczności leku oparta tylko na analizie jakościowej i ilościowej API jest niewystarczająca, ponieważ obecność w lekach innych substancji stanowiących zanieczyszczenia może wpływać na oczekiwany efekt farmakologiczny oraz znacznie zwiększyć toksyczność farmaceutyków. Na ogół obecność zanieczyszczeń w próbkach jest związana z brakiem kontroli warunków procesów produkcji, jak również ze stosowaniem substratów o niskiej jakości. Najczęściej stosowanym podejściem do

wykrywania zafałszowań leków jest wykorzystanie informacji na temat składu chemicznego analizowanych produktów farmaceutycznych. Metody stosowane do tego typu problemów badawczych polegają na analizie zawartości API oraz, jeśli to konieczne obejmują ocenę całościowego składu chemicznego próbki. Uzyskana zawartość substancji aktywnej jest następnie porównywana z ilością deklarowaną przez producenta danego leku. Jednakże w większości przypadków, oznaczenie API jest niewystarczające do potwierdzenia autentyczności leku. Alternatywne podejście zakłada, że próbki produktów farmaceutycznych są opisywane przez różnego rodzaju sygnały instrumentalne bez konieczności określenia ich składu chemicznego. Otrzymane sygnały instrumentalne są traktowane jako chemiczne odciski palca badanych próbek. Z definicji takie dane są wielowymiarowe co powoduje, że ich badanie i modelowanie wymaga użycia metod chemometrycznych w celu wyekstrahowania użytecznej informacji. Do charakteryzowania oraz weryfikacji autentyczności próbek leków stosuje się różne techniki instrumentalne [76]. Wśród nich znajdują się stosunkowo proste podejścia analityczne takie jak metody kolorymetryczne [77] czy dynamiczna analiza termiczna [78] jak również bardziej zaawansowane metody wśród, których można wymienić wysokosprawną chromatografię cieczową (HPLC) [79], chromatografię gazową (GC) [80], elektroforezę kapilarną [81], spektroskopię Ramana [82] oraz spektroskopię NMR [83].

W przemyśle farmaceutycznym źródłem informacji o składzie chemicznym badanych leków są najczęściej metody chromatograficzne [76]. Obszar ich stosowania jest bardzo szeroki, ponieważ pozwalają one uzyskać rozdział różnych składników mieszanin, które następnie analizuje się pod kątem jakościowym i ilościowym. Prosta metoda chromatografii cienkowarstwowej (TLC) jest wykorzystywana w wielu laboratoriach do weryfikacji składu leków. Popularność metody TLC wynika przede wszystkim z niskich kosztów analizy, niewymagającego wyposażenia instrumentalnego oraz prostej interpretacji wyników analizy. Ze względu na prostotę TLC można znaleźć liczne przykłady zastosowania chromatografii cienkowarstwowej w kontekście badań autentyczności leków opisane w literaturze [84,85]. Jednak najbardziej popularną techniką stosowana do analizy produktów farmaceutycznych jest wysokosprawna chromatografia cieczowa (HPLC), która jest traktowana jako metoda referencyjna w analizie jakościowej i ilościowej składników leków i służy jako metoda odniesienia do walidowania dużej liczby technik analitycznych. Chromatografy HPLC mogą być wyposażone w różnego rodzaju detektory takie jak na przykład spektrometria mas (MS), detektor z matrycą diodową (DAD) oraz detektor rozpraszania światła przez odparowanie

(ELS), które pomagają zwiększyć czułość, dokładność i precyzję stosowanej metody analizy leku. Te cechy są wymagane zwłaszcza wtedy, gdy prowadzona analiza chromatograficzna dotyczy składników chemicznych obecnych w próbce w małych stężeniach (np. zanieczyszczenia). Chromatografia gazowa (GC) jest stosowana do analizy substancji lotnych, które są trwałe w wysokich temperaturach. Podobnie jak HPLC, GC jest techniką charakteryzującą się dokładnością i powtarzalnością, a jej czułość zależy m.in. od zastosowanego detektora, którym może być m.in. spektrometr mas lub detektor płomieniowo-jonizacyjny. Chromatografia gazowa jest stosowana do kontroli leków głównie w kontekście oznaczania pozostałości rozpuszczalników i/lub lotnych zanieczyszczeń [80,86]. Ponieważ niewiele leków zawiera składniki lotne liczba zastosowań techniki GC jest znacznie mniejsza niż HPLC czy TLC.

Techniki spektroskopowe stanowią grupę metod wykorzystujących różne zakresy promieniowania elektromagnetycznego. Są one równie często stosowane w kontroli jakości produktów farmaceutycznych co techniki chromatograficzne. W celu weryfikacji autentyczności leków do najczęściej stosowanych należą spektroskopia w podczerwieni (NIR, FT-IR, FTIR-ATR), spektroskopia Ramana oraz spektroskopia jądrowego rezonansu magnetycznego (NMR) [62,87,88]. Technika NIR pozwala na szybką analizę próbek bez lub z niewielkim przygotowaniem próbki. Dodatkowo jej dużą zaletą jest możliwość uzyskania sygnału instrumentalnego bezpośrednio przez materiały stanowiące opakowanie leku (szkło lub plastikowe blistry). Ponadto, analiza prowadzona z wykorzystaniem techniki NIR jest niedestrukcyjna i stosunkowo tania.

Spektroskopia Ramana jest uważana jako wszechstronna technika ze względu na możliwość prowadzenia analizy leków w postaci stałej, ciekłej jak i gazowej. Podobnie jak technika NIR pozwala ona na prowadzenie analizy przez powłoki stanowiące opakowania analizowanych leków. Spektroskopia Ramana jest wysoce selektywna. Umożliwia ona rozróżnianie i identyfikację związków chemicznych o bardzo podobnej strukturze. W połączeniu z metodami chemometrycznymi, spektroskopia Ramana jest wykorzystywana w szerokim zakresie badań dotyczących identyfikacji autentyczności leków [63].

Spektroskopia NMR jest często stosowana do analizy substancji czynnych w lekach [83], oznaczania składu leków włącznie z analizą zanieczyszczeń w nich zawartych [89] oraz do monitorowania procesów produkcji produktów farmaceutycznych [90]. Ponadto widma NMR zawierają informacje na temat struktury analizowanych związków co pozwala na

identyfikacje składników zawartych w analizowanym leku. Jednak czułość tej techniki nie wystarcza aby analizować składniki obecne w próbkach w niskich stężeniach.

Uogólniając, wszystkie wspomniane metody instrumentalne dostarczają dużej ilości danych analitycznych, a każda próbka jest charakteryzowana za pomocą setek, a nawet tysięcy punktów pomiarowych, co sprawia, że eksploracja tego typu danych, ich modelowanie i interpretacja są skomplikowane. W tym celu wykorzystywane są często metody chemometryczne, które ułatwiają poszukiwanie i modelowanie danych wielowymiarowych [91].

Narzędzia chemometryczne stosowane do badania autentyczności leków są dość wszechstronne i posiadają duży obszar zastosowań. Jednak w danym problemie analitycznym wymagane jest często podejście wieloczynnikowe. Oznacza to, że różnice między autentycznymi i zafałszowanymi próbkami leków mogą wynikać z zestawu parametrów fizyko-chemicznych. W zależności od wybranej techniki analitycznej i zdefiniowanego celu badań, dane służące do oceny autentyczności leków mogą zawierać informację na temat stężenia substancji czynnej API, obecności substancji pomocniczych leku oraz zawartości zanieczyszczeń. Autentyczność leków może być także oceniana na podstawie całych sygnałów analitycznych (chemicznych odcisków palca) czy sygnałów instrumentalnych stanowiących profile zanieczyszczeń badanych próbek.

W zależności od typu analizowanych danych wymagana jest inna procedura ich wstępnego przygotowania do analizy chemometrycznej. W tej kwestii dla zbioru sygnałów analitycznych stosowane jest inne podejście niż dla zestawu zmiennych zawierających np. wartości stężeń wybranych komponentów próbek. Należy jednak zaznaczyć, że każda zastosowana metoda wstępnego przygotowania danych wpływa na wyniki analizy chemometrycznej, a tym samym na końcowe wnioski. Celem wstępnego przygotowania danych jest poprawa jakości wielowymiarowych sygnałów poprzez korygowanie i/lub niwelowanie wpływu niepożądanych efektów, co prowadzi do eliminacji niepożądanej zmienności. Jest to uzyskiwane poprzez zastosowanie różnego rodzaju transformacji matematycznych. Niepożądane źródła zmienności danych (szum, linia podstawowa, przesunięcia pików) powodują zaburzenia informacji chemicznej w nich zawartej oraz sprawiają, że jej wydobycie jest bardziej skomplikowane, a czasami nawet niemożliwe. Na ogół odpowiednio zastosowane metody wstępnego przygotowania danych pozwalają usunąć dużą część wariancji niezwiązanej z oczekiwanym efektem. Dlatego dalsza analiza skorygowanych danych i ich modelowanie będzie bardziej skuteczne i ułatwi interpretację

uzyskanych wyników. Niektóre metody wstępnego przygotowania danych są stosowane do odpowiedniego typu sygnałów instrumentalnych takich jak sygnały NIR czy sygnały chromatograficzne. W zależności od podjętego problemu badawczego, techniki wstępnego przygotowania danych mogą być stosowane dla indywidualnych zmiennych objaśniających lub dla poszczególnych próbek [92].

Wstępnie przygotowane dane opisujące problem zafałszowań leków są zazwyczaj analizowane za pomocą metod eksploracyjnych [93,94]. Pomagają one zobrazować ukrytą strukturę danych wielowymiarowych (zależności pomiędzy próbkami/parametrami). Metody grupowania danych dostarczają informacji o korelacji zmiennych pozwalając zidentyfikować zmienne, które nie posiadają dodatkowej informacji na temat analizowanych próbek dodatkowo umożliwiając identyfikację próbek stanowiących obiekty odległe. Eksploracja wielowymiarowych danych jest ważnym krokiem w analizie autentyczności leków, gdyż pozwala na identyfikację zależności pomiędzy próbkami zafałszowanymi i autentycznymi co w konsekwencji pozwala na odpowiedni dobór metod modelowania danych stosowanych w dalszych badaniach.

Kolejna grupą narzędzi chemometrycznych wykorzystywanych do wykrywania zafałszowań leków są metody uczenia z nadzorem. Są one stosowane do konstrukcji modeli diagnostycznych, która jest prowadzona z wykorzystaniem zmiennej zależnej definiującej przynależność próbek do analizowanych grup lub zawierającej informacje na temat badanej właściwości, na przykład stężenia substancji czynnej w analizowanych lekach. Metody dyskryminacyjne i klasyfikacyjne należące do metod uczenia z nadzorem są wykorzystywane do konstrukcji reguł logicznych, które pomagają odróżnić autentyczne i zafałszowane próbki leków w oparciu o różnego rodzaju informacje dotyczące ich składu chemicznego.

Główną różnicą pomiędzy podejściem klasyfikacyjnym i dyskryminacyjnym jest mechanizm, określający sposób w jaki analizowane próbki są przypisywane do istniejących grup. Metody klasyfikacyjne pozwalają na konstruowanie tzw. miękkich zasad klasyfikacji, które pozwalają, żeby próbka została przypisana do jednej grupy lub do większej liczby grup jednocześnie. Metody dyskryminacyjne wyznaczają tzw. twarde reguły klasyfikacji zgodnie z którymi próbka jest zawsze przyporządkowana tylko do jednej z analizowanych grup. Wśród metod dyskryminacyjnych najczęściej stosowanych w kontekście badania autentyczności leków wyróżnia się m.in. liniową analizę dyskryminacyjną (LDA), dyskryminacyjny wariant metody częściowych najmniejszych kwadratów (PLS-DA) oraz drzewa klasyfikacji i regresji (CART) [95,96]. Natomiast metody klasyfikacyjne stosowane

do tego typu badań to na przykład metoda modelowania indywidualnych grup (SIMCA), metoda klasyfikacji i analizy wpływu macierzy (M-CAIMAN) oraz metoda częściowych najmniejszych kwadratów z modelowaniem gęstości (PLS-DM) [97–99].

Dobór metody analitycznej oraz odpowiedniego podejścia chemometrycznego jest ściśle uzależniony od podjętego problemu badawczego (rodzaj badanego leku, stężenia analizowanych substancji itp.). Dlatego też trudne jest sformułowanie ogólnych wytycznych dotyczących poprawności stosowania narzędzi chemometrycznych do oceny autentyczności leków. Racjonalne podejście do wstępnego przygotowania danych i ich modelowania zakłada stosowanie możliwie najprostszych metod oraz modeli o możliwie najniższej kompleksowości. Wybór pomiędzy zastosowaniem metod dyskryminacyjnych i klasyfikacyjnych do oceny autentyczności leków powinien być racjonalnie przemyślany [100]. Metody klasyfikacji powinny być stosowane, gdy zmienność zafałszowanych próbek nie może być wystarczająco opisana. W tym przypadku wykorzystanie dwuklasowego modelu dyskryminacyjnego może prowadzić do błędnego rozpoznania nowych próbek jeśli wszystkie istotne źródła zmienności nie zostały ujęte podczas budowy modelu dyskryminacyjnego. Metody dyskryminacyjne mogą być stosowane, gdy zmienność danych dotyczących zafałszowanych danych jest ograniczona, na przykład, kiedy ilość zanieczyszczeń znajdująca się w zafałszowanych próbkach leku jest zawsze większa niż w autentycznych próbkach leku [101].

Więcej szczegółów na temat metod wykorzystywanych do badania autentyczności leków znajduje się w publikacji „Chemometrics and identification of counterfeit medicines - a review”, Journal of Pharmaceutical and Biomedical Analysis, 127 (2016) 112-122, która stanowi Załącznik nr 4 do niniejszej rozprawy doktorskiej.

5. Podsumowanie i wnioski

Uzyskane wyniki badań dowodzą, że metody chemometryczne znacznie wspierają proces identyfikacji i dyskryminacji próbek bazując na sygnałach instrumentalnych jakimi są chromatograficzne odciski palca. Odpowiednio dobrane metody chemometryczne dają możliwość eliminacji niepożądanych komponentów sygnałów instrumentalnych (szumu, linii podstawowej, przesunięć pików), które mogą negatywnie wpływać na uzyskane wyniki analizy danych chromatograficznych oraz ich interpretację. Zastosowane techniki chromatograficzne doskonale nadają się do rejestracji sygnałów instrumentalnych stanowiących chemiczne odciski palca ze względu na proporcje pomiędzy liczbą pików a liczbą substancji (jeden pik odpowiada jednemu związkowi chemicznemu). Wszystkie te cechy świadczą o unikalności chromatograficznych odcisków palca w kontekście całościowego opisu składu badanych próbek. Proponowane w pracy rozwiązania oparte na zaawansowanym modelowaniu chromatograficznych odcisków palca z wykorzystaniem narzędzi chemometrycznych pozwalają na eliminację konieczności przeprowadzania analizy jakościowej. Natomiast zastosowanie metod wyboru zmiennych istotnych dla określonego problemu dyskryminacyjnego może nie tylko skrócić czas rozdziału chromatograficznego jak również wskazać obszary czasów elucji zawierających frakcje związków odpowiedzialnych za różnicowanie grup. Dokładna analiza i interpretacja obszarów sygnałów chromatograficznych, które zostały uznane za istotne daje możliwość poznania związków stanowiących potencjalne markery opisanych procesów fałszowania lub identyfikacji substancji niebezpiecznych w wodzie.

Walidacja modeli wymaga zwrócenia uwagi na kilka parametrów walidacyjnych jednocześnie. Ponadto, niezwykle pomocne jest poznanie rozkładu ich wartości w funkcji kompleksowości modelu np. za pomocą proponowanej przeze mnie procedury Monte Carlo. Podejście to uwzględnia zmienność analizowanych zbiorów poprzez wielokrotne losowanie zbioru modelowego oraz zbiorów testowych, co z kolei wpływa na dokładniejszą ocenę optymalnej kompleksowości konstruowanego modelu dyskryminacyjnego.

Skonstruowane modele diagnostyczne, wykorzystane do dyskryminacji zafałszowanych próbek paliwa oraz leków, jak również próbek wody ze względu na zawartość substancji niebezpiecznych pozwoliły z bardzo dobrą efektywnością określić przynależność badanych próbek do określonych grup na podstawie całych sygnałów chromatograficznych lub bazując na wybranych zmiennych istotnych.

Wszystkie proponowane strategie analiz, które zostały opisane w niniejszej rozprawie doktorskiej, mogą być także implementowane w obszarach, w których problem badawczy dotyczy wykrywania zafałszowań lub oceny zawartości wybranych analitów za pomocą sygnałów chromatograficznych lub innego typu chemicznych odcisków palca.

6. Literatura

[1] G. Zadora, Laundering of “Illegal” Fuels -- a Forensic Chemistry Perspective, Acta Chimica Slovenica. 54 (2007) 110–113.

[2] P.-Y. Sacré, E. Deconinck, M. Daszykowski, P. Courselle, R. Vancauwenberghe, P. Chiap, J. Crommen, J.O. De Beer, Impurity fingerprints for the identification of counterfeit medicines—A feasibility study, Analytica Chimica Acta. 701 (2011) 224–231. doi:10.1016/j.aca.2011.05.041.

[3] E. Abdelaal, H.M. Ziena, M.M. Youssef, Adulteration of honey with high-fructose corn syrup: Detection by different methods, Food Chemistry. 48 (1993) 209–212. doi:10.1016/0308-8146(93)90061-J.

[4] P.L. Pisano, M.F. Silva, A.C. Olivieri, Anthocyanins as markers for the classification of Argentinean wines according to botanical and geographical origin. Chemometric modeling of liquid chromatography-mass spectrometry data, Food Chemistry. 175 (2015) 174–180. doi:10.1016/j.foodchem.2014.11.124.

[5] M. Daszykowski, B. Walczak, Use and abuse of chemometrics in chromatography, TrAC Trends in Analytical Chemistry. 25 (2006) 1081–1096. doi:10.1016/j.trac.2006.09.001.

[6] C. Tistaert, B. Dejaegher, Y. Vander Heyden, Chromatographic separation techniques and data handling methods for herbal fingerprints: a review, Analytica Chimica Acta. 690 (2011) 148–161. doi:10.1016/j.aca.2011.02.023.

[7] S.D. Brown, R. Tauler, B. Walczak, Data preprocessing, in: Comprehensive Chemometrics, Elsevier, London, 2009.

[8] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra, Applied Spectroscopy, AS. 43 (1989) 772–777.

[9] D.F. Thekkundan, S.C. Rutan, Denoising and signal-to-noise ratio enhancement: classical filtering, in: Comprehensive Chemometrics, Elsevier, Oxford, 2009: pp. 9–24.

[10] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of chemometrics and qualimetrics: part B, Elsevier, Amsterdam, 1998.

[11] Smoothing and Differentiation of Data by Simplified Least Squares Procedures. -Analytical Chemistry (ACS Publications). http://pubs.acs.org/doi/abs/10.1021/ac60214a047 (dostęp 14 marca 2016).

[12] B. Walczak, Wavelets in Chemistry, Elsevier, Amsterdam, 2000.

[13] P.H.C. Eilers, A perfect smoother, Analytical Chemistry. 75 (2003) 3631–3636. doi:10.1021/ac034173t.

[14] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, Journal of Chromatography A. 805 (1998) 17–35. doi:10.1016/S0021-9673(98)00021-1.

[15] M. Daszykowski, B. Walczak, Target selection for alignment of chromatographic signals obtained using monochannel detectors, Journal of Chromatography A. 1176 (2007) 1–11. doi:10.1016/j.chroma.2007.10.099.

[16] M. Daszykowski, Y. Vander Heyden, C. Boucon, B. Walczak, Automated alignment of one-dimensional chromatographic fingerprints, Journal of Chromatography A. 1217 (2010) 6127–6133. doi:10.1016/j.chroma.2010.08.008.

[17] B. Walczak, W. Wu, Fuzzy warping of chromatograms, Chemometrics and Intelligent Laboratory Systems. 77 (2005) 173–180. doi:10.1016/j.chemolab.2004.07.012.

[18] M. Daszykowski, B. Walczak, D.L. Massart, Projection methods in chemistry, Chemometrics and Intelligent Laboratory Systems. 65 (2003) 97–112. doi:10.1016/S0169-7439(02)00107-7.
[19] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and Intelligent Laboratory Systems. 2 (1987) 37–52.
[20] R.A. Fisher, The Use of Multiple Measurements in Taxonomic Problems, Annals of Eugenics. 7 (1936) 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.
[21] L. Breiman, J.H. Friedman, R.A. Olshen, C.G. Stone, Classification and regression trees, Wadsworth International Group. 93 (1984) 101.
[22] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory. 13 (1967) 21–27. doi:10.1109/TIT.1967.1053964.
[23] M. Barker, W. Rayens, Partial least squares for discrimination, Journal of Chemometrics. 17 (2003) 166–173. doi:10.1002/cem.785.
[24] T. Næs, T. Isaksson, T. Fearn, T. Davies, Mutivariate Calibration and Classification, NIR Publication, Chichester, 2002.
[25] T. Næs, H. Martens, Mutivariate Calibration, John Wiley & Sons, Chichester, 1989.
[26] L. Eriksson, E. Johansson, C. Wikström, Mixture design—design generation, PLS analysis, and model usage, Chemometrics and Intelligent Laboratory Systems. 43 (1998) 1–24. doi:10.1016/S0169-7439(98)00126-9.
[27] R.D. Snee, Validation of Regression Models: Methods and Examples, Technometrics. 19 (1977) 415–428. doi:10.1080/00401706.1977.10489581.
[28] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics. 11 (1969) 137–148. doi:10.1080/00401706.1969.10490666.
[29] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away, Journal of Chemometrics. 28 (2014) 213–225. doi:10.1002/cem.2609.
[30] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics. 44 (1988) 837–845.
[31] R. Wehrens, H. Putter, L.M.C. Buydens, The bootstrap: a tutorial, Chemometrics and Intelligent Laboratory Systems. 54 (2000) 35–52. doi:10.1016/S0169-7439(00)00102-7.
[32] Q.-S. Xu, Y.-Z. Liang, Monte Carlo cross validation, Chemometrics and Intelligent Laboratory Systems. 56 (2001) 1–11. doi:10.1016/S0169-7439(00)00122-2.
[33] Y.-Z.L. Qing-Song Xu, Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. Journal of Chemometrics, 18, 112-120, Journal of Chemometrics. 18 (2004) 112–120. doi:10.1002/cem.858.
[34] T.-T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recognition. 48 (2015) 2839–2846. doi:10.1016/j.patcog.2015.03.009.
[35] B. Efron, C. Stein, The Jackknife Estimate of Variance, Annals of Statistic. 9 (1981) 586–596. doi:10.1214/aos/1176345462.
[36] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, Chemometrics and Intelligent Laboratory Systems. 118 (2012) 62–69. doi:10.1016/j.chemolab.2012.07.010.
[37] C.M. Andersen, R. Bro, Variable selection in regression—a tutorial, Journal of Chemometrics. 24 (2010) 728–737. doi:10.1002/cem.1360.
[38] O.M. Kvalheim, T.V. Karstang, Interpretation of latent-variable regression models, Chemometrics and Intelligent Laboratory Systems. 7 (1989) 39–51. doi:10.1016/0169-7439(89)80110-8.
[39] T. Rajalahti, R. Arneberg, F.S. Berven, K.-M. Myhr, R.J. Ulvik, O.M. Kvalheim, Biomarker discovery in mass spectral profiles by means of selectivity ratio plot,

Chemometrics and Intelligent Laboratory Systems. 95 (2009) 35–48. doi:10.1016/j.chemolab.2008.08.004.
[40] V. Centner, D.-L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of Uninformative Variables for Multivariate Calibration, Analytical Chemistry. 68 (1996) 3851–3858. doi:10.1021/ac960321m.
[41] T.N. Tran, N.L. Afanador, L.M.C. Buydens, L. Blanchet, Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC), Chemometrics and Intelligent Laboratory Systems. 138 (2014) 153–160. doi:10.1016/j.chemolab.2014.08.005.
[42] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Multiand megavariate data analysis. Principles and applications., Umetrics Academy, Umea, 2001.
[43] Y. Vander Heyden, Extracting information from chromatographic herbal fingerprints, LC-GC Europe. 21, 438–443.
[44] D. Custers, M. Canfyn, P. Courselle, J.O. De Beer, S. Apers, E. Deconinck, Headspace–gas chromatographic fingerprints to discriminate and classify counterfeit medicines, Talanta. 123 (2014) 78–88. doi:10.1016/j.talanta.2014.01.020.
[45] M. Goodarzi, P.J. Russell, Y. Vander Heyden, Similarity analyses of chromatographic herbal fingerprints: A review, Analytica Chimica Acta. 804 (2013) 16–28. doi:10.1016/j.aca.2013.09.017.
[46] J. Orzel, M. Daszykowski, M. Kazura, D. de Beer, E. Joubert, A.E. Schulze, T. Beelders, A. de Villiers, C.J. Malherbe, B. Walczak, Modeling of the total antioxidant capacity of rooibos (Aspalathus linearis) tea infusions from chromatographic fingerprints and identification of potential antioxidant markers, Journal of Chromatography A. 1366 (2014) 101–109. doi:10.1016/j.chroma.2014.09.030.
[47] I. Stanimirova, B. Üstün, T. Cajka, K. Riddelova, J. Hajslova, L.M.C. Buydens, B. Walczak, Tracing the geographical origin of honeys based on volatile compounds profiles assessment using pattern recognition techniques, Food Chemistry. 118 (2010) 171–176. doi:10.1016/j.foodchem.2009.04.079.
[48] A.J. Charlton, M.S. Wrobel, I. Stanimirova, M. Daszykowski, H.H. Grundy, B. Walczak, Multivariate discrimination of wines with respect to their grape varieties and vintages, European Food Research and Technology. 231 (2010) 733–743. doi:10.1007/s00217-010-1299-2.
[49] R.L. White, P.D. Wentzell, M.A. Beasy, D.S. Clark, D.W. Grund, Taxonomy of Amanita mushrooms by pattern recognition of amino acid chromatographic data, Analytica Chimica Acta. 277 (1993) 333–346. doi:10.1016/0003-2670(93)80446-R.
[50] N.P. Mncwangi, A.M. Viljoen, J. Zhao, I. Vermaak, W. Chen, I. Khan, What the devil is in your phytomedicine? Exploring species substitution in Harpagophytum through chemometric modeling of 1H-NMR and UHPLC-MS datasets, Phytochemistry. 106 (2014) 104–115. doi:10.1016/j.phytochem.2014.06.012.
[51] L.S.M. Wiedemann, L.A. d'Avila, D.A. Azevedo, Adulteration detection of Brazilian gasoline samples by statistical analysis, Fuel. 84 (2005) 467–473. doi:10.1016/j.fuel.2004.09.013.
[52] L.F.P. Brandão, J.W.B. Braga, P.A.Z. Suarez, Determination of vegetable oils and fats adulterants in diesel oil by high performance liquid chromatography and multivariate methods, Journal of Chromatography A. 1225 (2012) 150–157. doi:10.1016/j.chroma.2011.12.076.
[53] J. Orzel, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Identifying the illegal removal from diesel oil of certain chemical markers that designate excise duty, Fuel. 117, Part A (2014) 224–229. doi:10.1016/j.fuel.2013.09.029.

[54] Rozporządzenie Ministra Finansów z dnia 20 sierpnia 2010 r. w sprawie znakowania i barwienia wyrobów energetycznych (Dz. U. 2010, Nr 157, poz. 1054).
[55] J. Orzel, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, B. Walczak, Simultaneous determination of Solvent Yellow 124 and Solvent Red 19 in diesel oil using fluorescence spectroscopy and chemometrics, Talanta. 101 (2012) 78–84. doi:10.1016/j.talanta.2012.08.031.
[56] B. Krakowska, I. Stanimirova, J. Orzel, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Detection of discoloration in diesel fuel based on gas chromatographic fingerprints, Analytical and Bioanalytical Chemistry. 407 (2014) 1159–1170. doi:10.1007/s00216-014-8332-4.
[57] N.M. Faber, R. Rajkó, How to avoid over-fitting in multivariate calibration—The conventional validation approach and an alternative, Analytica Chimica Acta. 595 (2007) 98–106. doi:10.1016/j.aca.2007.05.030.
[58] F.M. Fernandez, D. Hostetler, K. Powell, H. Kaur, M.D. Green, D.C. Mildenhall, P.N. Newton, Poor quality drugs: grand challenges in high throughput detection, countrywide sampling, and forensics in developing countries, Analyst. 136 (2011) 3073–3082. doi:10.1039/c0an00627k.
[59] R. Kumar, Recent Applications of Analytical techniques for counterfeit drug analysis: A Review, International Journal of PharmTech Research. 6 (2014) 646–665.
[60] P. de Peinder, M.J. Vredenbregt, T. Visser, D. de Kaste, Detection of Lipitor® counterfeits: A comparison of NIR and Raman spectroscopy in combination with chemometrics, Journal of Pharmaceutical and Biomedical Analysis. 47 (2008) 688–694. doi:10.1016/j.jpba.2008.02.016.
[61] M.J. Anzanello, R.S. Ortiz, R. Limberger, K. Mariotti, A framework for selecting analytical techniques in profiling authentic and counterfeit Viagra and Cialis, Forensic Science International. 235 (2014) 1–7. doi:10.1016/j.forsciint.2013.12.005.
[62] U. Holzgrabe, M. Malet-Martino, Analytical challenges in drug counterfeiting and falsification-The NMR approach, Journal of Pharmaceutical and Biomedical Analysis. 55 (2011) 679–687. doi:10.1016/j.jpba.2010.12.017.
[63] K. Dégardin, Y. Roggo, F. Been, P. Margot, Detection and chemical profiling of medicine counterfeits by Raman spectroscopy and chemometrics, Analytica Chimica Acta. 705 (2011) 334–341. doi:10.1016/j.aca.2011.07.043.
[64] C.R. Jung, R.S. Ortiz, R. Limberger, P. Mayorga, A new methodology for detection of counterfeit Viagra® and Cialis® tablets by image processing and statistical analysis, Forensic Science International. 216 (2012) 92–96. doi:10.1016/j.forsciint.2011.09.002.
[65] B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles, Analyst. 141 (2016) 1060–1070. doi:10.1039/C5AN01656H.
[66] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, Journal of Chemometrics. 29 (2015) 528–536. doi:10.1002/cem.2736.
[67] Regulation (EC) No 782/2003 of the European Parliament and of the Council of 14 April 2003 on the prohibition of organotin compounds on ships.
[68] ROZPORZĄDZENIE KOMISJI (UE) Nr 196/2010 z dnia 9 marca 2010 r. zmieniające załącznik I do rozporządzenia Parlamentu Europejskiego i Rady (WE) nr 689/2008 dotyczącego wywozu i przywozu niebezpiecznych chemikaliów.
[69] M. Bravo, G. Lespes, I. De Gregori, H. Pinochet, M.P. Gautier, Determination of organotin compounds by headspace solid-phase microextraction-gas

chromatography-pulsed flame-photometric detection (HS-SPME-GC-PFPD), Analytical and Bioanalytical Chemistry. 383 (2005) 1082–1089. doi:10.1007/s00216-005-0131-5.
[70] M. Bravo M., L.F. Aguilar, W. Quiroz V., A.C. Olivieri, G.M. Escandar, Determination of tributyltin at parts-per-trillion levels in natural waters by second-order multivariate calibration and fluorescence spectroscopy, Microchemical Journal. 106 (2013) 95–101. doi:10.1016/j.microc.2012.05.013.
[71] M. Barker, W. Rayens, Partial least squares for discrimination, Journal of Chemometrics. 17 (2003) 166–173. doi:10.1002/cem.785.
[72] M. Daszykowski, M. Korzen, B. Krakowska, K. Fabianczyk, Expert system for monitoring the tributyltin content in inland water samples, Chemometrics and Intelligent Laboratory Systems. 149, Part A (2015) 123–131. doi:10.1016/j.chemolab.2015.10.008.
[73] P.-Y. Sacré, E. Deconinck, T. De Beer, P. Courselle, R. Vancauwenberghe, P. Chiap, J. Crommen, J.O. De Beer, Comparison and combination of spectroscopic techniques for the detection of counterfeit medicines, Journal of Pharmaceutical and Biomedical Analysis. 53 (2010) 445–453. doi:10.1016/j.jpba.2010.05.012.
[74] O.Y. Rodionova, L.P. Houmøller, A.L. Pomerantsev, P. Geladi, J. Burger, V.L. Dorofeyev, A.P. Arzamastsev, NIR spectrometry for counterfeit drug detection: A feasibility study, Analytica Chimica Acta. 549 (2005) 151–158. doi:10.1016/j.aca.2005.06.018.
[75] K. Dégardin, Y. Roggo, P. Margot, Understanding and fighting the medicine counterfeit market, Journal of Pharmaceutical and Biomedical Analysis. 87 (2014) 167–175. doi:10.1016/j.jpba.2013.01.009.
[76] R. Martino, M. Malet-Martino, V. Gilard, S. Balayssac, Counterfeit drugs: analytical techniques for their identification, Analytical and Bioanalytical Chemistry. 398 (2010) 77–92. doi:10.1007/s00216-010-3748-y.
[77] M.T. Koesdjojo, Y. Wu, A. Boonloed, E.M. Dunfield, V.T. Remcho, Low-cost, high-speed identification of counterfeit antimalarial drugs on paper, Talanta. 130 (2014) 122-127. doi:10.1016/j.talanta.2014.05.050.
[78] S. Wilczyński, The use of dynamic thermal analysis to distinguish between genuine and counterfeit drugs, International Journal of Pharmaceutics. 490 (2015) 16–21. doi:10.1016/j.ijpharm.2015.04.077.
[79] A. Panusa, G. Multari, G. Incarnato, L. Gagliardi, High-performance liquid chromatography analysis of anti-inflammatory pharmaceuticals with ultraviolet and electrospray-mass spectrometry detection in suspected counterfeit homeopathic medicinal products, Journal of Pharmaceutical and Biomedical Analysis. 43 (2007) 1221–1227. doi:10.1016/j.jpba.2006.10.012.
[80] E. Deconinck, M. Canfyn, P.-Y. Sacré, S. Baudewyns, P. Courselle, J.O. De Beer, A validated GC–MS method for the determination and quantification of residual solvents in counterfeit tablets and capsules, Journal of Pharmaceutical and Biomedical Analysis. 70 (2012) 64–70. doi:10.1016/j.jpba.2012.05.022.
[81] R.D. Marini, E. Rozet, M.L.A. Montes, C. Rohrbasser, S. Roht, D. Rhème, P. Bonnabry, J. Schappler, J.-L. Veuthey, P. Hubert, S. Rudaz, Reliable low-cost capillary electrophoresis device for drug quality control and counterfeit medicines, Journal of Pharmaceutical and Biomedical Analysis. 53 (2010) 1278–1287. doi:10.1016/j.jpba.2010.07.026.
[82] C. Ricci, L. Nyadong, F. Yang, F.M. Fernandez, C.D. Brown, P.N. Newton, S.G. Kazarian, Assessment of hand-held Raman instrumentation for in situ screening for potentially counterfeit artesunate antimalarial tablets by FT-Raman spectroscopy and direct ionization mass spectrometry, Analytica Chimica Acta. 623 (2008) 178–186. doi:10.1016/j.aca.2008.06.007.

[83] V. Silvestre, V.M. Mboula, C. Jouitteau, S. Akoka, R.J. Robins, G.S. Remaud, Isotopic 13C NMR spectrometry to assess counterfeiting of active pharmaceutical ingredients: Site-specific 13C content of aspirin and paracetamol, Journal of Pharmaceutical and Biomedical Analysis. 50 (2009) 336–341. doi:10.1016/j.jpba.2009.04.030.
[84] Lukasz Komsta, M. Waksmundzka-Hajnos, J. Sherma, Thin Layer Chromatography in Drug Analysis, CRC Press, 2013.
[85] M.E. ElTantawy, L.I. Bebawy, R.F. Shokry, Chromatographic determination of clopidogrel bisulfate; detection and quantification of counterfeit Plavix® tablets, Bulletin of Faculty of Pharmacy, Cairo University. 52 (2014) 91–101. doi:10.1016/j.bfopcu.2014.04.003.
[86] D.B. da Justa Neves, R.G.A. Marcheti, E.D. Caldas, Incidence of anabolic steroid counterfeiting in Brazil, Forensic Science International. 228 (2013) 81–83. doi:10.1016/j.forsciint.2013.02.035.
[87] J. Luypaert, D.L. Massart, Y. Vander Heyden, Near-infrared spectroscopy applications in pharmaceutical analysis, Talanta. 72 (2007) 865–883. doi:10.1016/j.talanta.2006.12.023.
[88] S. Wartewig, R.H.H. Neubert, Pharmaceutical applications of Mid-IR and Raman spectroscopy, Advanced Drug Delivery Reviews. 57 (2005) 1144–1170. doi:10.1016/j.addr.2005.01.022.
[89] I. McEwen, A. Elmsjö, A. Lehnström, B. Hakkarainen, M. Johansson, Screening of counterfeit corticosteroid in creams and ointments by NMR spectroscopy, Journal of Pharmaceutical and Biomedical Analysis. 70 (2012) 245–250. doi:10.1016/j.jpba.2012.07.005.
[90] U. Holzgrabe, R. Deubner, C. Schollmayer, B. Waibel, Quantitative NMR spectroscopy—Applications in drug analysis, Journal of Pharmaceutical and Biomedical Analysis. 38 (2005) 806–812. doi:10.1016/j.jpba.2005.01.050.
[91] J. Djuris, Computer-Aided Applications in Pharmaceutical Technology, Elsevier, 2013.
[92] B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, Chemometrics and the identification of counterfeit medicines—A review, Journal of Pharmaceutical and Biomedical Analysis. 127 (2016) 112–122. doi:10.1016/j.jpba.2016.04.016.
[93] M. Daszykowski, B. Walczak, D.L. Massart, Projection methods in chemistry, Chemometrics and Intelligent Laboratory Systems. 65 (2003) 97–112. doi:10.1016/S0169-7439(02)00107-7.
[94] D.L. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Krieger Publishing Company, Florida, 1989.
[95] M.J. Anzanello, R.S. Ortiz, R.P. Limbergerb, P. Mayorga, A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes, Journal of Pharmaceutical and Biomedical Analysis. 83 (2013) 209–214. doi:10.1016/j.jpba.2013.05.004.
[96] E. Deconinck, P.Y. Sacré, D. Coomans, J. De Beer, Classification trees based on infrared spectroscopic data to discriminate between genuine and counterfeit medicines, Journal of Pharmaceutical and Biomedical Analysis. 57 (2012) 68–75. doi:10.1016/j.jpba.2011.08.036.
[97] M. Forina, C. Armanino, R. Leardi, G. Drava, A class-modelling technique based on potential functions, Journal of Chemometrics. 5 (1991) 435–453.
[98] M. Forina, P. Oliveri, S. Lanteri, M. Casale, Class-modeling techniques, classic and new, for old and new problems, Chemometrics and Intelligent Laboratory Systems. 93 (2008) 132–148. doi:10.1016/j.chemolab.2008.05.003.

[99] M.P. Derde, D.L. Massart, UNEQ: a disjoint modelling technique for pattern recognition based on normal distribution, Analytica Chimica Acta. 184 (1986) 33–51. doi:10.1016/S0003-2670(00)86468-5.
[100] O.Y. Rodionova, A.V. Titova, A.L. Pomerantsev, Discriminant analysis is an inappropriate method of authentication, TrAC Trends in Analytical Chemistry. 78 (2016) 17–22. doi:10.1016/j.trac.2016.01.010.
[101] B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles, Analyst. 141 (2016) 1060–1070. doi:10.1039/C5AN01656H.

7. Curriculum Vitae

Barbara Krakowska

data urodzenia: 13.11.1988 r

telefon: 505 023 188

miejsce zamieszkania: ul. Fliegera 14/19, Katowice

e-mail: bkrakowska@us.edu.pl

Wykształcenie

2012 –	Uniwersytet Śląski w Katowicach, Wydział Matematyki, Fizyki i Chemii, Instytut Chemii, studia doktoranckie
2011-2013	Uniwersytet Śląski w Katowicach, Wydział Pedagogiki i Psychologii, Podyplomowe Kwalifikacyjne Studia Pedagogiczne
2010 – 2012	Uniwersytet Śląski w Katowicach, Wydział Matematyki, Fizyki i Chemii, Instytut Chemii, studia II-go stopnia na kierunku chemia w zakresie chemii podstawowej
2007 – 2010	Uniwersytet Śląski w Katowicach, Wydział Matematyki, Fizyki i Chemii, Instytut Chemii, studia I-go stopnia na kierunku chemia w zakresie chemii podstawowej
2004 – 2007	IV Liceum Ogólnokształcące im. Hanki Sawickiej w Kielcach, klasa o profilu biologiczno – chemicznym

---

Zainteresowania naukowe

- Metody wstępnego przygotowania sygnałów instrumentalnych
- Analiza chromatograficznych odcisków palca
- Metody chemometryczne wykorzystywane do oceny autentyczności/jakości wybranych produktów
- Metody wyboru istotnych zmiennych
- Walidacja modeli diagnostycznych

Doświadczenie

- Staż w laboratorium Izby Celnej w Białej Podlaskiej (8.10.2015 – 30.10.2015) dotyczący analizy zafałszowań oleju napędowego
- Staż w laboratorium firmy Polcargo International w Szczecinie (26.01.2015 – 6.02.2015) dotyczący monitorowania substancji priorytetowych w próbkach środowiskowych
- Zajęcia laboratoryjne z chemii analitycznej, dla studentów I roku pierwszego stopnia na kierunku Chemia (2012/2013, 2013/2014, 2014/2015, 2015/2016 r.)
- Zajęcia laboratoryjne z programowania w środowisku Matlab, dla studentów II roku pierwszego stopnia, na kierunku Chemia (2013/2014 r.)
- Praktyki zawodowe w Gimnazjum nr 2 w Jędrzejowie (09.2012 r.)
- Praktyki zawodowe w laboratorium kontroli jakości w firmie Gomar Pińczów (09.2009)

Języki obce

- Język angielski – średniozaawansowany w mowie i piśmie
- Język norweski – podstawowy w mowie i piśmie

8. Dorobek naukowy

Publikacje opublikowane w czasopismach z listy filadelfijskiej wchodzące w cykl rozprawy doktorskiej

1. **B. Krakowska**, D. Custers, E. Deconinck, M. Daszykowski, Chemometrics and identification of counterfeit medicines – a review, Journal of Pharmaceutical and Biomedical Analysis, 127 (2016) 112-122; IF = 3,169; 35pkt.*
2. **B. Krakowska**, D. Custers, E. Deconinck, M. Daszykowski, The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles, Analyst, 141 (2016) 1060-1070; IF = 4,033; 40 pkt.*
3. M. Daszykowski, M. Korzeń, **B. Krakowska**, K. Fabiańczyk, Expert system for monitoring the tributyltin content in inland water samples, Chemometrics and Intelligent Laboratory Systems, 149 (2015) 123-131; IF = 2,217; 40 pkt*
4. **B. Krakowska**, I. Stanimirova, J. Orzeł, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Detection of discoloration in diesel fuel based on gas chromatographic fingerprints, Analytical and Bioanalytical Chemistry, 407 (2015) 1159-1170; IF = 3,125; 35pkt*

Sumaryczny IF = 12,544
Sumaryczna liczba punktów MNiSW* = 150

Publikacje opublikowane w czasopismach z listy filadelfijskiej

1. D. Custers, **B. Krakowska**, J.O. De Beer, P. Courselle, M. Daszykowski, S. Apers, E. Deconinck, Testing of complementarity of PDA and MS detectors using chromatographic fingerprinting of genuine and counterfeit samples containing sildenafil citrate, Analytical and Bioanalytical Chemistry, 408 (2016) 1643-1656; IF = 3,125; 35pkt.*
2. D. Custers, **B. Krakowska**, J.O. De Beer, P. Courselle, M. Daszykowski, S. Apers, E. Deconinck, Chromatographic impurity fingerprinting of genuine and counterfeit Cialis® as a means to compare the discriminating ability of PDA and MS detection, Talanta, 146 (2016) 540-548; IF = 4,035; 40pkt.*

3. R. Sitko, **B. Gliwińska**, B. Zawisza, B. Feist, Ultrasound-assisted solid-phase extraction using multiwalled carbon nanotubes for determination of cadmium by flame atomic absorption spectrometry, Journal of Analytical Atomic Spectrometry, 28 (2013) 405–410; IF = 3,396; 35pkt.*

Sumaryczny IF = 10,556

Sumaryczna liczba punktów MNiSW* = 110

* Punktacja zgodna z rokiem ukazania się publikacji według listy czasopism punktowanych MNiSW

Prezentacje ustne na konferencjach naukowych

1. D. Custers, B. Krakowska, P. Courselle, M. Daszykowski, S. Apers, E. Deconinck, Testing of complementarity of PDA and MS detectors using chromatographic fingerprinting of genuine and counterfeit Viagra®, XXXVIII Sympozjum nt. Chromatograficzne metody badania związków organicznych, Szczyrk (26-29 maja 2015)
2. B. Krakowska, J. Orzel, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Wykrywanie procederu fałszowania oleju napędowego przy użyciu metod chemometrycznych, Zjazd Wiosenny SSPTChem, 2014, Zawoja (9-13 kwietnia 2014)

Plakaty naukowe

1. B. Krakowska, J. Orzeł, I. Stanimirova, M. Sznajder, M. Zaleszczyk, I. Grabowski, M. Daszykowski, Studying the changes of excise duty components in diesel oil samples under influence of a reducing agent using gas chromatography with nitrogen chemiluminescence detector, XXXVIII Sympozjum nt. Chromatograficzne metody badania związków organicznych, Szczyrk (1-3 czerwca 2016)
2. B. Krakowska, J. Orzeł, A. Czarny, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Wykrywanie nielegalnego procederu odbarwiania oleju napędowego przy użyciu metod chemometrycznych, X Seminarium Naukowe „Aktualne Problemy Chemii Analitycznej, Katowice (15 maja 2016)

3. B. Krakowska, M. Daszykowski, Wykorzystanie metod chemometrycznych w kontroli autentyczności wybranych produktów, IV Ogólnopolska Konferencja „Pomiędzy naukami – Zjazd Fizyków i Chemików", Chorzów (18 września 2015)
4. B. Krakowska, M. Daszykowski, K. Fabiańczyk, M. Korzeń, Identification of tributyltin in environmental water samples supported by means of discriminant models, XXXVIII Sympozjum nt. Chromatograficzne metody badania związków organicznych, Szczyrk (26-29 maja 2015)
5. B. Krakowska, I. Stanimirova, D. Custers, E. Deconinck, M. Daszykowski, Wykorzystanie profilowania zanieczyszczeń do potwierdzenia autentyczności Viagry, IX Seminarium Naukowe „Aktualne Problemy Chemii Analitycznej, Katowice (15 maja 2015)
6. B. Krakowska, M. Daszykowski, Wykorzystanie metod chemometrycznych w kontekście kontroli jakości produktów farmaceutycznych, III Ogólnopolska Konferencja „Pomiędzy naukami – Zjazd Fizyków i Chemików", Chorzów (26 września 2014)
7. B. Krakowska, I. Stanimirova, D. Custers, E. Deconinck, M. Daszykowski, Detection of counterfeit medicines based on chromatographic impurity profiles, XXXVII Sympozjum nt. Chromatograficzne metody badania związków organicznych, Szczyrk (11-13 czerwca 2014)
8. B. Krakowska, I. Stanimirova, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Porównanie metod wyboru zmiennych na przykładzie ich zastosowania do budowy modelu dyskryminacyjnego, VIII Seminarium Naukowe „Aktualne Problemy Chemii Analitycznej, Katowice (16 maja 2014)
9. B. Krakowska, J. Orzeł, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Partial Least Squares for detection of diesel oil adulteration, PLS 2014, Paryż (19-21 maj 2014)
10. B. Krakowska, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Zastosowanie metod chemometrycznych do wykrywania procederu fałszowania oleju napędowego, Zjazd Zimowy SSPTChem, Łódź (7 grudnia 2013)
11. B. Krakowska, M. Daszykowski, Wykorzystanie metod chemometrycznych w analizie jakości wybranych produktów, II Ogólnopolska Konferencja „Pomiędzy naukami – Zjazd Fizyków i Chemików", Chorzów (27 września 2013)

12. B. Krakowska, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Comparative analysis of diesel oil samples of different origin based on chromatographic fingerprints, XXXVI Sympozjum nt. Chromatograficzne metody badania związków organicznych, Szczyrk (5 czerwca 2013)
13. B. Krakowska, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Wykrywanie procederu fałszowania oleju napędowego przy użyciu metod chemometrycznych, VII Seminarium Naukowe „Aktualne Problemy Chemii Analitycznej, Katowice (17 maja 2013)
14. B. Krakowska, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Eksploracja danych chromatograficznych opisujących próbki oleju napędowego, Zjazd Wiosenny SSPTChem, Augustów (11-18 kwietnia 2013)

15. B. Gliwińska (Krakowska), K. Kocot, B. Feist, R. Sitko, Zagęszczanie jonów miedzi i kadmu z wykorzystaniem wielościennych nanorurek weglowych techniką absorpcyjnej spektrometrii atomowej, VI Seminarium Naukowe „Aktualne Problemy Chemii Analitycznej", Katowice (18 maja 2012)

Załącznik 1

Publikacja:	Detection of discoloration in diesel fuel based on gas chromatographic fingerprints
Autorzy:	B. Krakowska I. Stanimirova-Daszykowska J. Orzeł M. Daszykowski I. Grabowski G. Zaleszczyk M. Sznajder
Czasopismo:	Analytical and Bioanalytical Chemistry
Wartość współczynnika Impact Factor	3,125

Anal Bioanal Chem (2015) 407:1159–1170
DOI 10.1007/s00216-014-8332-4

RESEARCH PAPER

Detection of discoloration in diesel fuel based on gas chromatographic fingerprints

Barbara Krakowska · Ivana Stanimirova · Joanna Orzel · Michal Daszykowski · Ireneusz Grabowski · Grzegorz Zaleszczyk · Mirosław Sznajder

Received: 13 August 2014 / Revised: 31 October 2014 / Accepted: 7 November 2014 / Published online: 19 November 2014

Abstract In the countries of the European Community, diesel fuel samples are spiked with Solvent Yellow 124 and either Solvent Red 19 or Solvent Red 164. Their presence at a given concentration indicates the specific tax rate and determines the usage of fuel. The removal of these so-called excise duty components, which is known as fuel "laundering", is an illegal action that causes a substantial loss in a government's budget. The aim of our study was to prove that genuine diesel fuel samples and their counterfeit variants (obtained from a simulated sorption process) can be differentiated by using their gas chromatographic fingerprints that are registered with a flame ionization detector. To achieve this aim, a discriminant partial least squares analysis, PLS-DA, for the genuine and counterfeit oil fingerprints after a baseline correction and the alignment of peaks was constructed and validated. Uninformative variables elimination (UVE), variable importance in projection (VIP), and selectivity ratio (SR), which were coupled with a bootstrap procedure, were adapted in PLS-DA in order to limit the possibility of model overfitting. Several major chemical components within the regions that are relevant to the discriminant problem were suggested as being the most influential. We also found that the bootstrap variants of UVE-PLS-DA and SR-PLS-DA have excellent predictive abilities for a limited number of gas chromatographic features, 14 and 16, respectively. This conclusion was also supported by the unitary values that were obtained for the area under the receiver operating curve (AUC) independently for the model and test sets.

B. Krakowska · I. Stanimirova · J. Orzel · M. Daszykowski (✉)
Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland
e-mail: mdaszyk@us.edu.pl

I. Grabowski · G. Zaleszczyk · M. Sznajder
Customs Chamber of Customs Laboratory in Biala Podlaska, 21 Celnikow Polskich Street, 21-500 Biala Podlaska, Poland

Keywords Excise duty components · Partial least squares discriminant analysis · Uninformative variable elimination-partial least squares · Variable selection · Fuel "laundering" · Bootstrapping

Introduction

The steadily increasing level in the consumption of petrol oil worldwide generates considerable profits for the petroleum industry and an increase in the price of petrol oil. Apart from the economic factors, the price of fuel is dependent on the local regulations that define the level of excise tax. In general, many countries apply different levels of excise duty on fuel depending on its usage. For instance, the diesel fuel that is used for heating purposes and in agricultural machinery in Poland has a rebated excise tax that is regulated by law in comparison to the diesel fuel that is used for regular transport. In order to differentiate an oil product with respect to its usage, a dye (red dye Solvent Red 19 or Solvent Red 164) and a marker (Solvent Yellow 124) are deliberately added [1]. The presence of these specific excise duty components neither modifies the physicochemical properties of fuel nor limits its further usage. The substantial financial reward that can be gained from the difference in excise tax has stimulated the illegal practice of removing the excise duty components from rebated fuel and selling it at a higher price. This procedure is known as fuel "laundering" and can be done by an adsorption process using widely available materials. The laundering of commercially available fuel causes a substantial loss in a government's budget and therefore, the development of analytical procedures for the detection of counterfeit diesel fuel is extremely necessary.

Detection of fuel laundering specifically requires an analytical technique that is capable of revealing chemical changes in the composition of the fuel, since the removal of the excise

duty components does not influence its physicochemical parameters. In our previous studies, we developed an analytical methodology to detect any chemical changes before and after a simulated laundering process by using diesel fuel fingerprints that were obtained using fluorescence spectroscopy [2, 3]. It was confirmed that genuine samples can definitely be discriminated from samples after the laundering process. However, only limited information about the chemical composition of complex mixtures can be obtained from their fluorescence fingerprints and that is why gas chromatography coupled with flame ionization detection (GC-FID) was investigated in this study. Gas chromatography (GC) is one of the most popular separation techniques for studying complex petrochemical samples because the chromatograms contain comprehensive chemical information. Gas chromatographic fingerprints [4] are widely used for monitoring quality and/or for identification purposes. These fingerprinting techniques have also been accepted by the World Health Organization (WHO) for the quality assessment of herbal products [5]. To the best of our knowledge, up to the present, GC-FID has not been used for studying the laundering process. The reason is that the excise duty components are not stable and degrade under a high temperature. Their instability was studied and confirmed throughout our experiment when using GC with the nitrogen chemiluminescence detector (sensitive to the presence of compounds containing nitrogen in their structures). Furthermore, the low concentration levels of excise duty components in diesel oil samples make the identification of their peaks among or under the peaks of major sample components difficult. The larger the number of peaks the harder the separation and quantification of analytes with the GC technique is, and this may provide to a failure in the characterization of excise duty components. That is why, the excise duty components are mainly determined with either spectroscopic or HPLC-based techniques. In fact, only the presence or absence of excise duty components is not indicative for a possible laundering process, but the GC-FID fingerprints may contain information about the overall chemical characteristics of samples before and after laundering.

In general, a comparative analysis of chromatographic fingerprints does not require the qualitative or quantitative evaluation of chemical components in samples, but advanced chemometric techniques [6] are required. On the one hand, the costs of the analysis are greatly reduced because no certified reference materials/standards are required and while on the other, important regions in the chromatographic fingerprints that are related to the phenomena being studied (e.g., discrimination/classification of two or more groups of samples) can be found using well-validated multivariate chemometric methods. Once the important regions of chromatographic fingerprints are identified using a variable selection method [7], the corresponding fractions can be collected and further analyzed in detail using an orthogonal chromatographic system or a complementary analytical method. A methodology that combines the fingerprint approach and chemometric analysis has gained popularity in many fields of science and technology in recent years including those such as the development of a method for the estimation of the total antioxidant capacity of green tea [8], the comparative analysis of extraction performance under different conditions [9], the analysis of secondary metabolites in citrus fruits peels [10], classification of petroleum products [11], etc.

In order to investigate whether it is possible to detect diesel fuel laundering, the excise duty components of a number of samples that were obtained from different suppliers in Poland were removed using an adsorption process. Genuine diesel fuel samples and their counterfeit variants were analyzed using gas chromatography coupled with flame ionization detector. Differences between these two groups of samples were studied using the partial least squares discriminant analysis, PLS-DA [12]. The removal of baseline and the alignment of peaks were performed to the sample chromatographic fingerprints using penalized asymmetric least squares approach (PAsLS) [1], and correlation optimized warping (COW) [2], respectively. In order to identify the key regions that are related to the chemical differences of sample groups, the PLS-DA approach was extended with variable selection. In this study, uninformative variable elimination-partial least squares discriminant analysis (UVE-PLS-DA) [13], PLS-DA combined with variable importance in projection (VIP) [14], and selectivity ratio (SR) were investigated [15, 16]. The effect of variable selection was monitored using a bootstrap procedure and the area under the receiver operating curve (AUC) and the sensitivity, specificity, and efficiency for the independent test set were adopted as figures of merit as well.

Experimental

A total of 31 samples of diesel fuel were collected from different fuel suppliers located in Poland in accordance with the sampling requirements that are specified in the PN-EN ISO 3170:2004 norm. The samples covered the majority of diesel fuel sources that are available for regular customers, and crude oil used for production had a different geographical origin (Poland, Belarus, Lithuania, and The Netherlands). Each investigated sample fulfilled the norm specifications, and thus, could be considered to be representative for a given batch of diesel fuel. Prior to further analysis, the samples were stored at room temperature (ca. 20 °C).

Registration of the GC-FID fingerprints

The samples of diesel fuel were analyzed twice using a gas chromatographic system (Agilent Technologies 6890N) equipped with a flame ionization detector before and after

the removal of the excise duty components. Separation of the components of the mixture was performed using a RTX-5 Restek column, 60 m × 0.25 mm i.d. and 0.25 μm film thickness with helium as carrier gas (1.3 ml min^{-1} constant flow rate, gas purity 5.0). The following temperature program was used: initial temperature 50 °C raised up to 320 °C by 3 °C per minute; total analysis time 100 min. Other settings of applied chromatographic method are as follows: injection mode split (split ratio 20:1); injection temperature 300 °C; injection volume 0.1 μL.

Processing (laundering) of diesel fuel samples

Every diesel fuel sample was subjected to a specific laboratory treatment that was aimed at removing the excise duty components (the dye and marker). The following laboratory procedure was applied to each genuine diesel fuel sample. Ten milliliters of a sample was placed in a plastic test tube (15 ml) with 2 % of the adsorbent and shaken vigorously for 5 s. Each test tube was shaken two to three times within a 30-min period. Afterwards, each sample was centrifuged at 3500 rpms and the supernatant that was obtained was analyzed as is described in the "Registration of the GC-FID fingerprints" section.

Theory

Preprocessing of chromatographic fingerprints

Instrumental signals, e.g., chromatographic fingerprints, consist of three components that are expressed at different levels along the signal's domain. These are a baseline, a noise, and a pure analytical signal. Each component of the signal has a different frequency range. The noise component has the highest frequency due to the rapid changes within a small amplitude. The baseline component has a very smooth form with a low amplitude of changes and its frequency is the lowest. A pure analytical signal has an intermediate frequency as compared to the frequencies of the baseline and noise components.

Even though the chromatographic conditions are the subject of optimization, chromatographic fingerprints often contain a substantial baseline component. The baseline shape is often irreproducible due to various effects and thus, can influence the construction of multivariate models. It should effectively be removed in order to diminish the negative influence of an overpronounced and fluctuating baseline. To date, many methods have been proposed for removal of the baseline. Among them, the penalized asymmetric least squares method (PAsLS) has found numerous applications. This was the method of choice in our study because of its simplicity and efficiency. More details about the PAsLS method, including definition of its objective function and input parameters, can be found in ref. [17].

In addition to baseline correction, chromatographic fingerprints often require alignment to correct peak shifts. They are the result of different factors that influence the elution time, including the unavoidable effect of column aging. Peak shifts in different chromatograms that originate from the same substance strongly affect the further multivariate data analysis as well as modeling and therefore, their correction is mandatory. The correlation optimized warping approach (COW) is a standard technique that is used for the alignment of peaks [18]. Peak shifts are corrected by stretching and compressing corresponding sections in the target signal and the signal that is being aligned. This is done by maximizing the correlation coefficient between these two signals. In the course of the alignment procedure, the target signal serves as a template for matching chromatographic peaks of every signal from the whole set [19]. An extensive description of the COW method and selection of input parameters is provided in ref. [18].

Principal component analysis

Principal component analysis (PCA) is a bilinear projection method that is used to visualize and compress multivariate data [20, 21]. With this method, a collection of chromatographic fingerprints, which is organized into a data matrix, is represented as the product of the score and loading vectors, which are called the principal components. The principal components are found by maximizing the description of data variance. A display of score and loading vectors is usually presented for selected pairs of principal components. The proximity of the points on the score plot reflects the chemical similarities among the samples that are described by their chromatographic fingerprints. The loading values (weights) provide information about the relative importance of the variables (fraction(s) of the mixture that is eluted from a chromatographic column within a certain range of elution time) into the construction of a given principal component. Owing to its bilinear character, the score and loading vectors help in assessing any chemical similarities among samples and the loadings indicate the impact of parameters on data structure observed on score projections.

Partial least squares discriminant analysis

Partial least squares discriminant analysis (PLS-DA) is a variant of the classic partial least squares regression model that aims to discriminate groups of samples [22, 23]. The belongingness of a sample to a certain group is indicated using a categorical dependent variable, y. For a two-class discriminant problem, which is within the scope of this study, samples

of groups could be coded using a bipolar or a binary dependent vector with elements "−1" and "+1" or "0" and "1" [12].

The PLS-DA model is built using a balanced set of model set samples (the same number of samples from each group) that represent the possible sources of variance characteristic for the two groups of samples well [24]. It is important to emphasize that the selection of the model set samples is crucial for the future prediction of the properties of the PLS-DA model. A uniform scatter of samples over the experimental domain ensures that all sources of variability are taken into account when model is built. The Kennard and Stone or the Duplex algorithm [25] can be used for this uniform selection of the subset for each group of samples separately.

In order to construct the PLS-DA model, its complexity is optimized so that the covariance between a set of latent variables and the response variable, y, is a maximum. This is also a key step that has an impact on the future performance of the model. Different cross-validation procedures are frequently used [26] to assess the optimal number of latent PLS-DA variables. Their purpose is to obtain error estimates by perturbing models that are built with an increasing number of latent variables. In general, the cross-validation procedure is an iterative elimination of samples from a model set and an estimation of the prediction error for samples that are removed from the interim model. The final error estimates that are obtained from a series of models with a definite number of latent factors are averaged and displayed as a function of the number of latent factors. Usually, the performance of a discriminant model is presented with figures of merit that are based on the number of correctly recognized samples. Selected figures of merit such as area under the curve, sensitivity, and specificity are defined in the following section. However, the root mean square error of cross-validation is a more sensitive estimate with respect to model complexity and is also an indication of the spread of predicted values.

Performance of a discriminant model

There are many measures that can characterize the performance of a discriminant model. They are calculated independently for model and test samples. The most popular measure of a model's performance is the correct discrimination rate, which indicates the number of samples that are correctly recognized using a given discriminant model. Additional figures of merit such as sensitivity (also known as true positive rate, TPR) and specificity (true negative rate, TNR) are derived based on the number of true positive (TP) and true negative samples (TN) as well as false positive (FP) and false negative samples (FN). Sensitivity for a given group of samples is defined as the ratio of the number of true positive samples to the sum of true positive and false negative samples. Specificity expresses the ratio of the number of true negative samples to the sum of the true negative and false positive samples.

The receiver operating characteristic curve, better known as the ROC curve, is an alternative approach to score and illustrate the performance of a discriminant method. The ROC curve summarizes the performance of a discriminant model and displays the trade-off between the true positive rate and false positive rate (1−specificity) as a function of a model parameter. The convex shape of the ROC curve, i.e., above the line of the unit slope, is an indication of a superior model performance rather than a random guess. The closer the area under the curve (AUC) value is to one, the better the discrimination performance is. Therefore, the perfect discriminant model is characterized by a unitary AUC.

In order to obtain honest estimates of a model's performance, the bootstrap approach [27] can be adopted. The main idea of the bootstrap approach is to draw, in a random manner, an assumed number of samples from each group in order to form a model set that is used to construct a discriminant model. The remaining samples form the test set and help to estimate the accuracy of the prediction. Since the bootstrap procedure is repeated many times (hundreds of times), the distribution of figures of merit that are being considered is sampled. As a result, the mean value and standard deviation of any parameter that describes the accuracy of the model can be provided in order to illustrate its uncertainty based on multiple random selections of samples. An alternative approach is based on a permutation test. The permutation procedure aims to obtain a reference null distribution (discrimination results are expected to be insignificant) of a certain statistics that are generated from discriminant model that is constructed for the dependent categorical variable that reflects the random assignment of samples to existing groups [29].

Identification of the relevant explanatory variables for the PLS-DA model

In chemical studies, the number of variables often greatly exceeds the number of samples. That is why chemical data, including chromatographic fingerprints, contain many variables that are noisy and/or unreliable or redundant to the discrimination of the groups. It is known that such explanatory variables affect the prediction properties of the PLS estimator [28] and increase the complexity of a model. Reducing the complexity of a model and obtaining an easier model interpretation and a possible improvement in prediction can be achieved by the variable selection. In the context of PLS-DA, the large number of explanatory variables compared to the number of objects significantly increases the probability of good discrimination by chance. As was illustrated in [29], it is possible to obtain a perfect PLS-DA discrimination for randomly generated data with large ratio of variables to objects. Therefore, the

validation of any discriminant model is essential for its practical use. Many variable selection approaches are described in the literature [7]. Three main categories can be described—the filter, wrapper, and embedded methods [28].

Uninformative variable elimination-partial least squares

The UVE-PLS wrapper method was proposed to eliminate the uninformative explanatory variables that carry similar information to random variables [13]. In order to distinguish informative variables from uninformative, an experimental data matrix $X(m\times n)$ is augmented with a matrix of noisy variables, N. The noise matrix, N, contains the normally distributed random numbers of a small amplitude ($m\times n$*). These noisy variables that have a small variance and negligible covariance with the modeled response variable, in principle, do not influence the construction of the PLS-DA model. During the construction of the PLS-DA model, the stability of the jth variable (experimental and artificial), s_j, is evaluated based on the jackknifing procedure. The stability of the jth variable is defined as the ratio between the mean value and standard deviation of the regression coefficients, s_j, is:

$$s_j = \frac{b_j}{\mathrm{std}(\mathbf{b}_j)} \tag{1}$$

where, $\mathbf{b}_j$ is the vector that contains the regression coefficients of the jth variable that was obtained from jackknifing of the PLS-DA model.

Uninformative variables are identified as those with the lower absolute stability of the regression coefficients than the maximal absolute value of the stability of the regression coefficients observed for the noisy variables. Uninformative variables are eliminated from the data and the final model is constructed.

Variable importance in projection

Variable importance in projection (VIP) is a simple filter-based variable selection approach that is proposed to assess the relevance of the variables in the PLS-DA model [30, 31]. The importance of the jth variable is expressed by its VIP_j score, which is defined as:

$$\mathrm{VIP}_j = \sqrt{\frac{\sum_{f=1}^{F} w_{jf}^2 \cdot \mathrm{SSY}_f \cdot J}{\mathrm{SSY}_t \cdot F}} \tag{2}$$

where w_{jf} is the PLS weight value of the jth variable and the fth component, SSY_f is the sum of squares of the dependent variable that was obtained from the discriminant model with f (f=1, 2, …, F) components, J is the number of variables, SSY_t is the total sum of squares of the dependent variable, and F is the number of PLS components evaluated.

The importance of a variable is considered to be highly influential on the PLS model when its VIP score is above 1.0, moderately influential if the VIP score is within the range of 0.8 to 1.0 or if a variable has a small influence—its VIP is below 0.8 [32]. The procedure for the elimination of the variables below a given threshold can be repeated several times in order to reduce the number of variables.

Selectivity ratio

The selectivity ratio is another criterion that can be applied in order to filter out irrelevant variables [15, 16]. Irrelevant variables are considered to be those that are not related to the response variable even though they may have large variances. The larger the selectivity ratio of a variable, the greater the correlation with the response variable is. Once the PLS-DA model of a definite complexity is built, the so-called target projection transformation or target rotation is performed so that several PLS-DA components are represented by a target projection score ($m\times 1$) vector. Then, a target loading vector ($1\times n$) is obtained from the projection of a model set on to the normalized target score vector. Multiplying the target projection score and loading vectors gives a target projection matrix of dimensions ($m\times n$). Thus, the original matrix $\mathbf{X}$ of a model set can be represented as the sum of two matrices—a target projection matrix that contains the information about the PLS-DA model of definite complexity and a residual matrix. A quantitative measure of the selectivity of each variable for the discrimination of groups is the value of the ratio of the sum of squares of the target projection matrix to the residual sum of squares.

Bootstrap variants of variable selection methods

A bootstrap strategy was adopted to all of the three methods in order to estimate the effect of the variable selection procedure. The VIP-PLS-DA, SR-PLS-DA, and UVE-PLS-DA models were constructed and validated using an independent test set. The final PLS-DA models for the relevant variables were validated using an independent test set (these samples were not considered in construction of the model) and characterized by the average AUC values for the model set from 1000 bootstrap samples with a replacement and the AUC value for the test set.

For each bootstrap sample of UVE-PLS-DA, a model set containing the chromatographic signals that were selected (with a replacement) from each group was augmented with

100,000 noisy variables that were formed by random numbers drawn from normal distribution (multiplied by a constant factor $c=10^{-12}$). Relevant variables were then selected as those variables with absolute stabilities of their regression coefficients above a cut-off value corresponding to 99.9 % of the maximum value of the absolute stabilities of the regression coefficients for the noisy variables. In fact, 1000 bootstrap samples resulted in 1000 sets of the selected variables and therefore, the variable relevance to the final model is evaluated by the so-called selection frequency, which indicates the percentage of times a variable is selected in the model.

The bootstrap methodology using VIPs and SRs for variable elimination is similar [33]. Basically, it consists of two steps. Firstly, the PLS-DA model with the optimal complexity was constructed for each bootstrap sample and the VIP scores or SRs for the variables were obtained. Secondly, the irrelevant variables were identified as those for which their average values of VIP scores and SRs were below a selected cut-off value. In this study, a cut-off value of 0.8 was found to be optimal for both methods. The threshold value in SR-PLS-DA was selected using the discriminating variables test, DIVA test, and the selectivity ratio plot. In contrast to the SR-PLS-DA method, which was applied only once, the VIP-PLS-DA method was applied three times in a sequential manner in order to reduce the relatively high number of variables that are usually selected when it is only applied once [33].

Results and discussion

Preprocessing of the GC-FID fingerprints

Because chromatographic signals were collected at a high sampling rate that contained many measuring points, they were resampled using linear interpolation in order to simplify the further modeling. The initial sampling rate of 140,000 sampling points of the GC-FID fingerprints was reduced to 25,000 without a substantial loss of the quality of the signal. Figure 1a presents a typical example of chromatogram obtained from a complex diesel oil sample. Due to limitations of the chromatographic method, many peaks overlap and are not baseline separated. High and sharp chromatographic peaks represent the major components of a sample and are found at characteristic bulge of a baseline. In general, a large number of not fully resolved peaks characteristic for components at relatively low concentrations are found at the base line. Prior to the chemometric analysis, the baseline was removed using the PAsLS method. An acceptable baseline approximation was achieved for $\lambda=10{,}000$ and $p=0.001$.

A further detailed analysis of the GC-FID fingerprints collected also revealed a problem with the peak shifts.

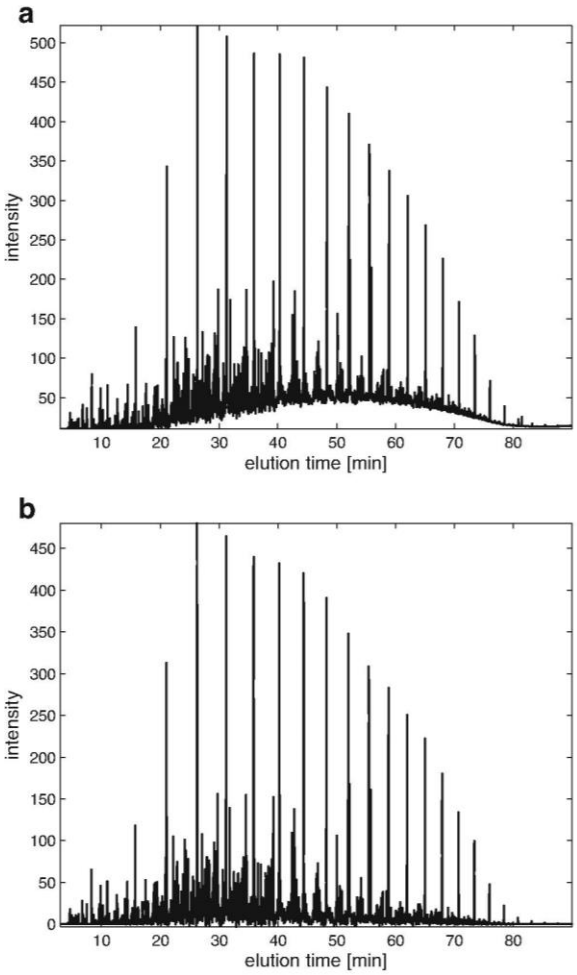


Fig. 1 Exemplary GC-FID fingerprint: **a** before and **b** after baseline removal

The correlation optimized warping method was used in order to correct the peak shifts. In this study, the reference chromatogram was selected as described in [19]. A different number of sections (starting with the length of a section corresponding to an average peak width of 50 sampling points) and values of the slack parameter were evaluated. A satisfactory alignment was achieved when the alignment was carried out for 250 sections (100 sampling points per section) and the slack parameter was equal to three for the majority of the GC-FID fingerprints. The signals after a baseline correction and alignment are presented in Fig. 1b.

The smallest value of the initial correlation coefficient between a signal and target was about 0.220, whereas the largest value was 0.989. A few fingerprints, which had relatively low correlation coefficients with respect to the target signal (compared to majority of signals), were aligned again

using different input parameters. In general, most of the GC-FID fingerprints were characterized by correlation coefficients that were higher than 0.8 after the alignment procedure. The smallest correlation coefficient that was observed was 0.804 and the largest value was 0.999. To illustrate the effect of the alignment procedure, histograms of the initial and the final (after alignment) correlation coefficients that were computed between each fingerprint and the target signal are presented in Fig. 2.

Preprocessed GC-FID fingerprints of genuine and counterfeit samples were further modeled using multivariate discriminant methods in order to verify the possibility of their discrimination.

Exploration of the GC-FID fingerprints

Potential differences in the chemical composition of diesel fuel samples were studied using the PCA method. Preprocessed GC-FID fingerprints (baseline corrected, aligned, and mean-centered) of genuine and counterfeit samples can be modeled with two principal components that describe 73.68 % of the total data variance. Projection of the samples onto the space that was defined by the first two principal components allows some conclusions about their chemical similarities to be drawn. Each point on the PC 1-PC 2 projection (Fig. 3a) represents one GC-FID fingerprint (sample). Genuine and discolored samples were denoted as "+" and "○", respectively. In Fig. 3b, for a better clarity of presentation pairs of samples authentic and counterfeit are connected with a line. Two groups of diesel fuel samples can be observed along the PC 1 axis and another two groups along the PC 2 axis. By analyzing the score projections in Fig. 3a, one can conclude that the laundering process itself is not substantially influential for the separation of the samples along PC 1 and PC 2. The corresponding loading plots in Fig. 3c, d, which show the ranges of the elution times, indicates the two chromatographic peaks that are responsible for the differences between the two groups of samples. These two peaks correspond to the mixtures that were eluted at ca. 65.21 and 65.24 min. They can be attributed to the methyl esters of fatty acids (FAME). The FAME compounds cannot be considered as possible markers for the laundering process. They are deliberately added to diesel oil during its production and are present in the studied samples regardless the laundering process. In Poland, any manufacturer is allowed to add up to 7 % (*V*/*V*) of FAME, but the differences in the total amount of FAME from batch to batch of diesel oil depend on the temporary production and economic requirements. The group of diesel oil samples characterized by positive score values along PC 1 (see Fig. 3a) contains FAME, the content of which varies in the range of 4.1 % (*V*/*V*) and 5.3 (*V*/*V*). Other variability sources such as different producers, origin of crude oil used in the production process,

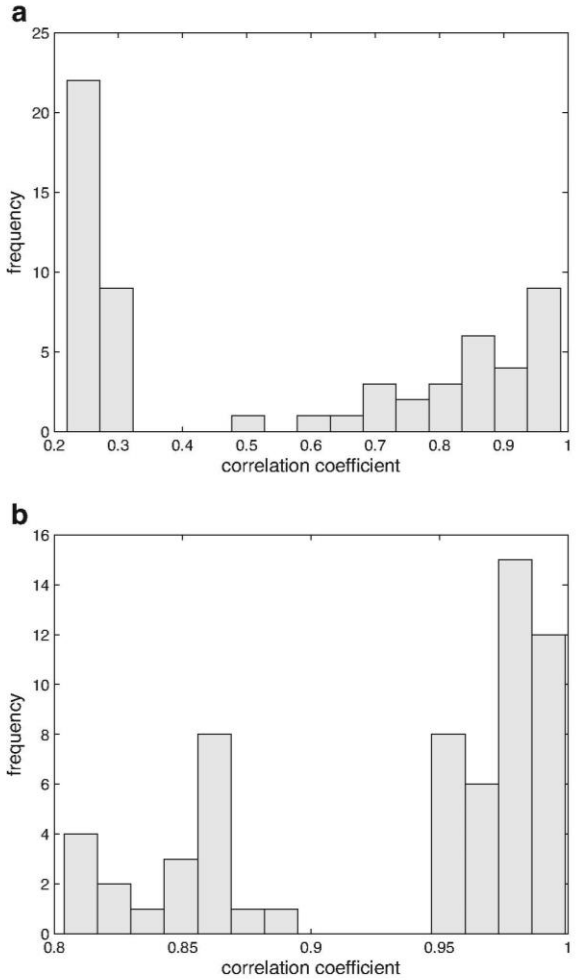


Fig. 2 Histograms of correlation coefficients calculated between each chromatographic fingerprint and a target signal: **a** before and **b** after alignment using COW

production process itself have an impact on forming clusters of samples.

Unfortunately, no groups of samples that underwent the laundering process were revealed in the other score projections built for consecutive pairs of the selected principal components. Therefore, the discrimination of groups along with the directions that describe the largest data variance in the experimental space is impossible. However, this does not necessarily mean that the supervised discrimination between the groups is also impossible. For this reason, the next step of the chemometric processing of GC-FID fingerprints was aimed at the construction of a supervised PLS-DA model to possibly support the differences in the chemical composition of genuine and counterfeit diesel fuel samples. On the other hand, Fig. 3a provides evidence that chemical composition of laundered

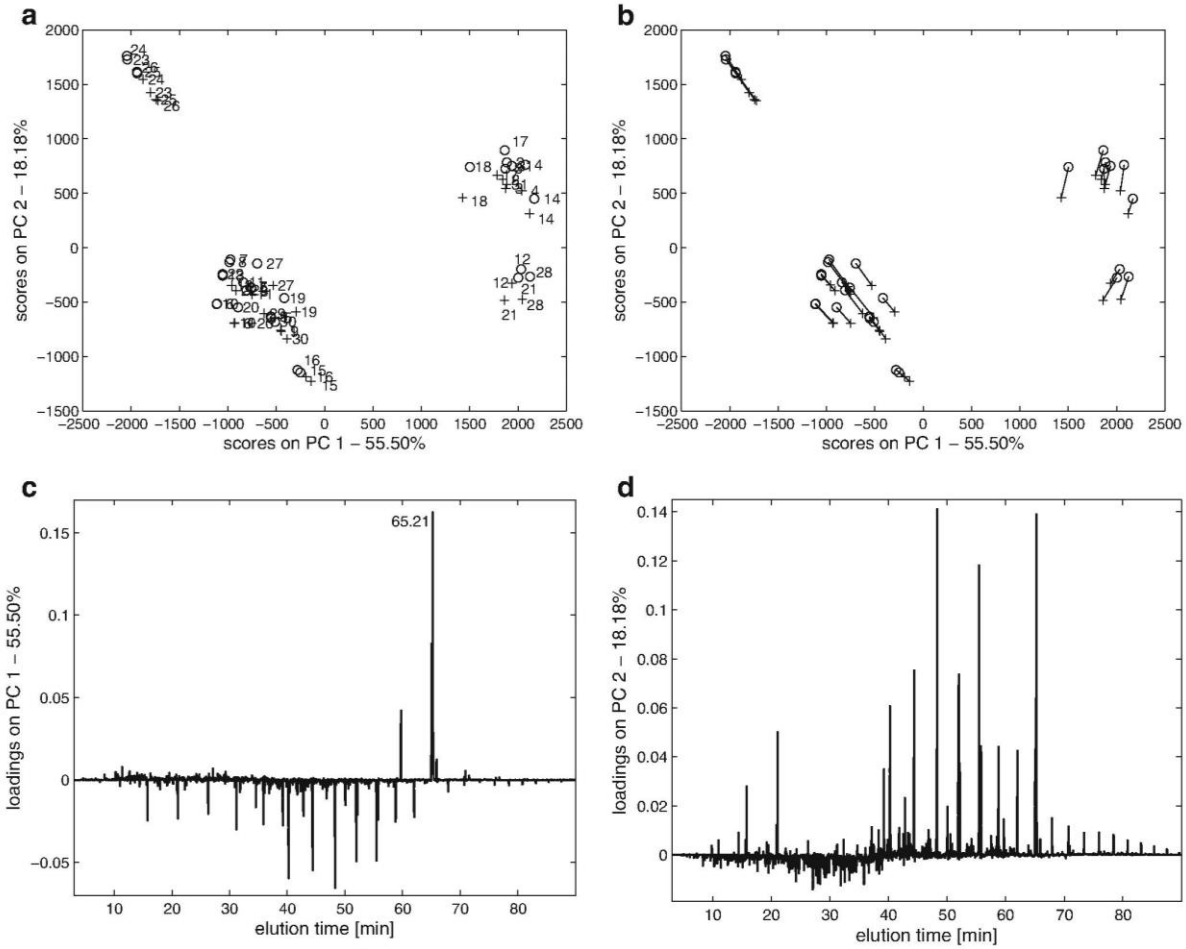


Fig. 3 Projection of samples (scores) onto space defined by first two principal components: **a** samples denoted as *plus sign* are authentic and samples denoted as *empty circle* are after the laundering process and **b** illustration of corresponding pairs of samples authentic *plus sign* and counterfeit variant *empty circle* (i.e. after the laundering process). Loadings as a function of retention time for: **c** PC 1 and **d** PC 2

samples is different from chemical composition of authentic ones. In Fig. 3a, corresponding pairs of samples (authentic samples marked as "+" and counterfeit samples marked as "○") are connected with a line. Samples after laundering are shifted with respect to its authentic variant. Since counterfeit samples are found in the same clusters, chemical composition is modified in a moderate degree and most probably concern components at low concentrations (minor components).

Construction of the PLS-DA model

Prior to the construction of the PLS-DA model, model diesel fuel samples were selected according to the following scheme. A total of 21 samples from the genuine group of diesel fuel samples were chosen using the Kennard and Stone algorithm in order to cover all of the possible sources of variability [25]. Genuine diesel fuel samples were coded as "+1" to reflect the presence of the excise duty components. The second group of the counterfeit diesel fuel samples contained the same samples, but after the laundering process. The samples of this group were coded as "−1" in order to indicate the absence of the excise duty components. The remaining ten diesel fuel samples and their ten counterfeit variants formed the test set and were used to characterize the predictive abilities of the model. The optimal number of latent factors, which were required to build a PLS-DA model for each bootstrap sample (selected with replacement from the model set), was selected based on the leave-one-out cross-validation procedure. As is indicated in Table 1, the PLS-DA model helps in discriminating all of the samples from the test set correctly. The excellent discrimination results that were obtained from the PLS-DA model

Table 1 Performance of partial least squares discriminant models with and without the variable selection scheme

Type of model	No. of variables	*f*	AUC model set	AUC test set	Sensitivity, specificity model set (%)	Sensitivity, specificity test set (%)
PLS-DA	25,000	6	1.000	1.000	100.00 100.00	100.00 100.00
UVE-PLS-DA	14	9	1.000	1.000	90.00 100.00	100.00 100.00
VIP-PLS-DA	265	6	0.996	0.970	95.24 95.24	90.00 90.00
SR-PLS-DA	16	3	1.000	1.000	100.00 100.00	100.00 100.00

support the hypothesis that the process of diesel fuel laundering can be detected based on the diesel fuel GC-FID fingerprints.

The bootstrap variable selection methods were then considered in order to avoid the possibility of presenting results of an overfitted discriminant model. Figure 4 illustrates the number of relevant variables that were selected from UVE-PLS-DA as a function of the percentage of the variable selection frequency.

In 96 % of all of the bootstrap subsets (1000 subsets drawn with replacement), only 14 variables out of 24,966 were found to be the most relevant. They corresponded to the mixtures that were eluted from the chromatographic column after 23.937, 23.940, 23.944, 24.786, 24.789, 24.793, 24.796, 25.398, 25.402, 27.556, 27.559, 27.563, 40.881, and 40.884 min. A list of potential chemical components, eluted from a column at selected retention times, is provided in Table 2. Identification of the chromatographic peaks described in Tables 2 and 3 was based on retention time index obtained from GC-MS (GC Agilent Technologies 7890A with MS 5975C detector) and supported by the NIST 2011 library. Compared to the corresponding genuine fuel samples, these components were found at lower levels in the laundered fuel samples.

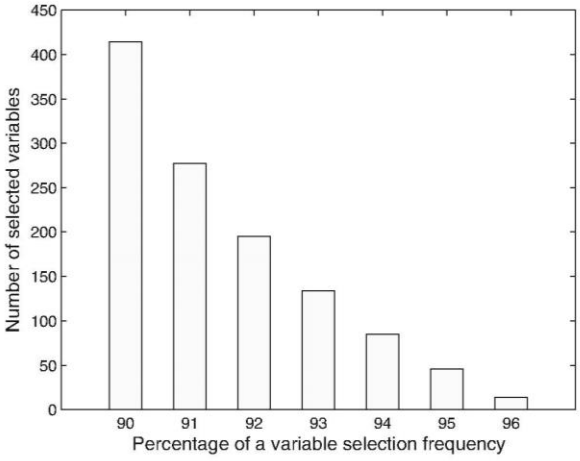


Fig. 4 The number of relevant variables identified using the UVE-PLS-DA approach as a function of percentage of the variable selection frequency

The final UVE-PLS-DA model was built and validated with an independent test set using the selected variables. The bootstrap procedure with a replacement was again carried out in order to evaluate the effect of the variable selection. The UVE-PLS-DA model was characterized by a unitary average AUC value for the model and test set. Only one sample from the model set was recognized incorrectly as a genuine sample. This results in a sensitivity of a 90 % and a highest specificity of 100 %. The sensitivity, specificity, accuracy, and correct classification rate for the final discriminant model with nine PLS factors and for the test set samples are presented in Table 1.

The final discriminant model with six latent factors that was built for 265 variables that were selected with the bootstrap VIP-PLS-DA procedure presented average AUC values of 0.996 and 0.970 for the model and test set, respectively. One sample in each group of the model and test set was incorrectly recognized using the final PLS-DA model. Both figures of merit (sensitivity and specificity) for the model set were equal to 95.24 %, whereas they were at a level of 90.00 % for the test set samples.

As was mentioned earlier, the average selectivity ratio for each variable was obtained from the bootstrap procedure. Irrelevant variables were identified as those with an absolute average selectivity ratio below the threshold value of 0.8. Only 16 variables were recognized as relevant. They correspond to the mixtures that were eluted at ca. 7.052, 7.055, 23.982, 23.985, 23.989, 23.991, 23.996, 23.999, 32.292, 32.296, 32.299, 32.891, 32.894, 32.898, 38.508, and 47.058 min (see Table 3).

The optimal PLS-DA model that was constructed for 16 variables had only three latent factors and offered an excellent discrimination performance for the model and test set samples. Once again, the unitary average AUC values for the model

Table 2 Identification of chemical compounds found in mixtures eluted at retention times indicated as relevant using the UVE-PLS-DA approach

Peak number	Retention time [min]	Possible compound
1	23.937	Benzene, 1-methyl-3-propyl formula: $C_{10}H_{14}$
	23.940	
	23.944	
2	24.786	Benzene, 1-methyl-4-propyl formula: $C_{10}H_{14}$
	24.789	
	24.793	
	24.796	
3	25.398	Benzene, 1-ethyl-2,4-dimethyl formula: $C_{10}H_{14}$
	25.402	
4	27.556	Benzene, 1,2,3,5-tetramethyl formula: $C_{10}H_{14}$
	27.559	
	27.563	
5	40.881	*n*-Paraffin C_{14}
	40.884	

and test set as well as sensitivities and specificities of 100 % were obtained.

A comparison of the results that were obtained from variable selection approaches

In general, it is not easy to decide which variable selection method will show the best performance for a given discriminant problem. A number of variable selection methods for PLS-DA have been developed for this purpose. Here, we considered two filter methods and one wrapper method. UVE-PLS-DA and SR-PLS-DA are methods that are specially designed to select variables that correlate with the response variable, even though they can have low variances. Both methods achieve this through the different mechanisms that were described earlier. In this study, they presented the best predictive abilities (sensitivity and specificity of 100 % for the test set) and a small number of selected variables in comparison with the VIP-PLS-DA. In our study, even though a sequential scheme of variable selection was adopted in VIP-PLS-DA, a large number of variables were found to be important. These variables are in fact the olefin substances that are present in the highest concentrations in diesel fuel and definitely present large absolute sizes (have large variances in the PLS-DA model), but from a chemical point of view, they are not necessarily related to the laundering process. Moreover, the VIP-PLS-DA method showed the worst predictive

Table 3 Identification of chemical compounds found in mixtures eluted at retention times indicated as relevant using the SR-PLS-DA approach (NI—not identified)

Peak number	Retention time [min]	Possible compound
1	7.052	NI
	7.055	
2	23.982	4-Ethyloheptan formula: C_9H_{20} or 1-octanol, 2-butyl formula: $C_{12}H_{26}$
	23.985	
	23.989	
	23.991	
	23.996	
	23.999	
3	32.292	Phytol formula: $C_{20}H_{40}O$
	32.296	
	32.299	
4	32.891	Compounds containing oxygen, e.g., 1-propene, 2-nitro-3-(1-cyclooctenyl) formula: $C_{11}H_{17}NO_2$
	32.894	
	32.898	
5	38.508	NI
6	47.058	Pentadecane, 3-methyl formula: $C_{16}H_{34}$

performance among the three variable selection discriminant methods. The other two methods, UVE-PLS-DA and SR-PLS-DA, found two sets of important variables. A closer look of these sets of variables indicates that the variables that were selected by SR-PLS-DA have a clear chemical interpretation since these are polar substances, the concentrations of which decrease during the adsorption laundering process.

Conclusions

We came to several important conclusions in this work. The data exploration that was performed using PCA did not show any characteristic distribution of chromatographic fingerprints of diesel fuel samples with respect to the differences in the sample contents. On the other hand, all three variable selection methods, UVE-PLS-DA, SR-PLS-DA, and VIP-PLS-DA, show potential in the detection of laundering process. Among them, VIP-PLS-DA presented the worst predictive performance and the largest number of selected variables. The other two methods showed a sensitivity, specificity, and efficiency of 100 % for the test set using a small number of variables (14 and 16). A closer look at the variables that were selected by both methods indicated that the variables that were obtained from SR-PLS-DA have a straightforward chemical interpretation. These are polar substances, the concentration of which decreases during the adsorption laundering process. Therefore, it appears that PLS-DA with the variables that were selected using their selectivity ratios is the method of choice for the detection of illegal diesel fuel discoloration. Although a larger number of commercially available diesel fuel samples should be considered in order to definitely determine the general use of the methodology, these results indicate the potential and practical use of proposed method.

Acknowledgments BK and JO are grateful for the financial support within the framework of the DoktoRIS program—scholarship program for innovative Silesia co-financed by the European Union under the European Social Fund.

Conflict of interest We wish to confirm that there are no known conflicts of interest associated with this publication.

References

1. Disposition of Polish Minister of Finance concerning marking and dying of energy products, (Dz. U. 2010, No. 157, item 1054)
2. Orzel J, Daszykowski M, Grabowski I, Zaleszczyk G, Sznajder M, Walczak B (2012) Simultaneous determination of solvent yellow 124 and solvent red 19 using fluorescence spectroscopy and chemometrics. Talanta 101:78–84
3. Orzel J, Daszykowski M, Grabowski I, Zaleszczyk G, Sznajder M (2013) Identifying the illegal removal from diesel oil of certain chemical markers that designate excise duty. Fuel 117:224–229
4. Vander Heyden Y (2008) Extracting information from chromatographic herbal fingerprints. LC-GC Europe 21:438–443
5. Tistaert C, Dejaegher B, Chataigné G, Rivière C, Nguyen Hoai N, Chau Van M, Quetin-Leclercq J, Vander Heyden Y (2012) Potential antioxidant compounds in Mallotus species fingerprints. Part II: fingerprint alignment, data analysis and peak identification. Anal Chim Acta 721:35–43
6. Daszykowski M, Walczak B (2006) Use and abuse of chemometrics in chromatography. Trends Anal Chem 25:1081–1096
7. Guyon I, Gunn S, Nikravesh M, Zadeh LA (2006) Feature extraction: foundations and applications. Springer, Berlin
8. van Nederkassel AM, Daszykowski M, Massart DL, Vander Heyden Y (2005) Prediction of total green tea antioxidant capacity from chromatograms by multivariate modeling. J Chromatogr A 1096: 177–186
9. Faghihi J, Jiang X, Vierling R, Goldman S, Sharfstein S, Sarver J, Erhardt P (2001) Reproducibility of the high-performance liquid chromatographic fingerprints obtained from two soybean cultivars and a selected progeny. J Chromatogr A 915:61–74
10. Parastar M, Jalali-Heravi M, Sereshti H, Mani-Varnosfaderani A (2012) Chromatographic fingerprint analysis of secondary metabolites in citrus fruits peels using gas chromatography–mass spectrometry combined with advanced chemometric methods. J Chromatogr A 1251:176–187
11. Nielsen NJ, Ballabio D, Tomasi G, Todeschini R, Christensen JH (2012) Chemometric analysis of gas chromatography with flame ionisation detection chromatograms: a novel method for classification of petroleum products. J Chromatogr A 1238:121–127
12. Barker M, Rayens W (2003) Partial least squares for discrimination. J Chemometr 17:166–173
13. Centner V, Massart DL, Noord OE, Jong S, Vandeginste BM, Sterna C (1996) Elimination of uninformative variables for multivariate calibration. Anal Chem 68:3851–3858
14. Andersen CM, Bro R (2010) Variable selection in regression—a tutorial. J Chemometr 24:728–737
15. Kvalheim OM, Karstang TV (1989) Interpretation of latent-variable regression models. Chemom Intell Lab Syst 7:39–51
16. Rajalahti T, Arneberg R, Berven F, Myhr KM, Ulvik RJ, Kvalheim OM (2009) Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemom Intell Lab Syst 95:35–48
17. Eilers PHC (2003) A perfect smoother. Anal Chem 75:3631–3636
18. Nielsen N, Carstensen J, Smedsgaard J (1998) Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. J Chromatogr A 805:17–35
19. Daszykowski M, Walczak B (2007) Target selection for alignment of chromatographic signals obtained using monochannel detectors. J Chromatogr A 1176:1–11
20. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemom Intell Lab Syst 2:37–52
21. Daszykowski M, Walczak B, Massart DL (2003) Projection methods in chemistry. Chemom Intell Lab Syst 65:97–112
22. Næs T, Isaksson T, Fearn T, Davies T (2002) Multivariate calibration and classification. NIR, Chichester
23. Martens H, Næs T (1989) Multivariate calibration. Wiley, Chichester
24. Brereton RG, Lloyd GR (2014) Partial least squares discriminant analysis: taking the magic away. J Chemometr 28:213–225
25. Daszykowski M, Walczak B, Massart DL (2002) Representative subset selection. Anal Chim Acta 468:91–103

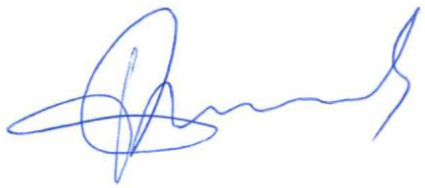
26. Xu QS, Liang YZ (2001) Monte Carlo cross validation. Chemom Intell Lab Syst 56:1–11
27. Wehrens R, Putter H, Buydens LMC (2000) The bootstrap: a tutorial. Chemom Intell Lab Syst 54:35–52
28. Mehmood T, Liland KH, Snipen L, Sæbø S (2012) A review of variable selection methods in partial least squares regression. Chemom Intell Lab Syst 118:62–69
29. Westerhuis JA, Hoefsloot HCJ, Smit S, Vis DJ, Smilde AK, van Velzen EJJ, van Duijnhoven JPM, van Dorsten FA (2008) Assessment of PLSDA cross validation. Metabolomics 4:81–89
30. Favilla S, Durante C, Li Vigni M, Cocchi M (2013) Assessing feature relevance in NPLS models by VIP. Chemom Intell Lab Syst 129:76–86
31. Wold S, Johansson E, Cocchi M (1993) 3D SAR in drug design; theory, method and applications. Escom, Leiden, pp 523–550
32. Eriksson L, Johansson E, Kettaneh-Wold N, Wold S (2001) Multi- and megavariate data analysis. Principles and applications. Umetrics Academy, Umea
33. Bro R, Nielsen HJ, Savorani F, Kjeldahl K, Christensen IJ, Brünner N, Lawaetz AJ (2013) Data fusion in metabolomic cancer diagnostics. Metabolomics 9:3–8

dr Ivana Stanimirova-Daszykowska
Instytut Chemii
Uniwersytet Śląski
ul. Szkolna 9
40-006 Katowice

Katowice 29.06.2016

Oświadczam, że w artykule pt. „Detection of discoloration in diesel fuel based on gas chromatographic fingerprints" opublikowanym w czasopiśmie Analytical and Bioanalytical Chemistry, 407 (2015) 1159-1170 mój udział polegał na:

- współpracy w dokonaniu wyboru zmiennych z zastosowaniem metod UVE, SR i VIP,
- dyskusji uzyskanych wyników,
- pomocy w tworzeniu manuskryptu.



dr Joanna Orzeł
Instytut Chemii
Uniwersytet Śląski
ul. Szkolna 9
40-006 Katowice

Katowice 23.06.2016

Oświadczam, że w artykule pt. „Detection of discoloration in diesel fuel based on gas chromatographic fingerprints" opublikowanym w czasopiśmie Analytical and Bioanalytical Chemistry, 407 (2015) 1159-1170 mój udział polegał na:

- pomocy w dokonaniu wstępnej organizacji danych,
- weryfikacji poprawnego wykorzystywania narzędzi chemometrycznych (PCA, PLS-DA),
- pomocy w interpretacji uzyskanych wyników.

Joanna Orzeł

dr hab. Michał Daszykowski, prof. UŚ
Instytut Chemii
Uniwersytet Śląski
ul. Szkolna 9
40-006 Katowice

Katowice 29.06.2016

Oświadczam, że w artykule pt. „Detection of discoloration in diesel fuel based on gas chromatographic fingerprints" opublikowanym w czasopiśmie Analytical and Bioanalytical Chemistry, 407 (2015) 1159-1170 mój udział polegał na:

- współtworzeniu hipotezy badawczej i ogólnej koncepcji badań,
- weryfikacji poprawnego wykorzystywania narzędzi chemometrycznych (PAsLS, COW, PLS-DA, SR, VIP, UVE),
- pomocy w interpretacji wyników uzyskanych przez doktorantkę,
- opiece i merytorycznym nadzorze procesu przygotowania manuskryptu,
- pomocy w przeprowadzeniu procedury redakcyjnej i przygotowaniu odpowiedzi na recenzje,
- dokonaniu ostatecznej korekty artykułu.

Michał Daszykowski

dr Ireneusz Grabowski
Laboratorium Celne
Izba Celna w Białej Podlaskiej
Terminal Samochodowy w Koroszczynie
21-550 Terespol

Biała Podlaska, 24.06.2016

Oświadczam, że w artykule pt. „Detection of discoloration in diesel fuel based on gas chromatographic fingerprints" opublikowanym w czasopiśmie Analytical and Bioanalytical Chemistry, 407 (2015) 1159-1170 mój udział polegał na:

- pomocy w postawieniu hipotezy badawczej i ustaleniu ogólnej koncepcji badań,
- pomocy w interpretacji uzyskanych wyników.

mgr Grzegorz Zaleszczyk
Laboratorium Celne
Izba Celna w Białej Podlaskiej
Terminal Samochodowy w Koroszczynie
21-550 Terespol

Biała Podlaska, 24.06.2016

Oświadczam, że w artykule pt. „Detection of discoloration in diesel fuel based on gas chromatographic fingerprints" opublikowanym w czasopiśmie Analytical and Bioanalytical Chemistry, 407 (2015) 1159-1170 mój udział polegał na:

- przeprowadzeniu procesu odbarwiania oleju napędowego.

Grzegorz Zaleszczyk

mgr inż. Mirosław Sznajder
Laboratorium Celne
Izba Celna w Białej Podlaskiej
Terminal Samochodowy w Koroszczynie
21-550 Terespol

Biała Podlaska, 24.06.2016

Oświadczam, że w artykule pt. „Detection of discoloration in diesel fuel based on gas chromatographic fingerprints" opublikowanym w czasopiśmie Analytical and Bioanalytical Chemistry, 407 (2015) 1159-1170 mój udział polegał na:

- uzyskaniu sygnałów chromatograficznych analizowanych próbek.

Mirosław Sznajder

Załącznik 2

Publikacja:	The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles
Autorzy:	B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski
Czasopismo:	Analyst
Wartość współczynnika Impact Factor	4,033

Analyst

PAPER

View Article Online
View Journal | View Issue

Cite this: *Analyst*, 2016, **141**, 1060

The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles

B. Krakowska,[a] D. Custers,[b,c] E. Deconinck[b] and M. Daszykowski*[a]

The aim of this work was to develop a general framework for the validation of discriminant models based on the Monte Carlo approach that is used in the context of authenticity studies based on chromatographic impurity profiles. The performance of the validation approach was applied to evaluate the usefulness of the diagnostic logic rule obtained from the partial least squares discriminant model (PLS-DA) that was built to discriminate authentic Viagra® samples from counterfeits (a two-class problem). The major advantage of the proposed validation framework stems from the possibility of obtaining distributions for different figures of merit that describe the PLS-DA model such as, *e.g.*, sensitivity, specificity, correct classification rate and area under the curve in a function of model complexity. Therefore, one can quickly evaluate their uncertainty estimates. Moreover, the Monte Carlo model validation allows balanced sets of training samples to be designed, which is required at the stage of the construction of PLS-DA and is recommended in order to obtain fair estimates that are based on an independent set of samples. In this study, as an illustrative example, 46 authentic Viagra® samples and 97 counterfeit samples were analyzed and described by their impurity profiles that were determined using high performance liquid chromatography with photodiode array detection and further discriminated using the PLS-DA approach. In addition, we demonstrated how to extend the Monte Carlo validation framework with four different variable selection schemes: the elimination of uninformative variables, the importance of a variable in projections, selectivity ratio and significance multivariate correlation. The best PLS-DA model was based on a subset of variables that were selected using the variable importance in the projection approach. For an independent test set, average estimates with the corresponding standard deviation (based on 1000 Monte Carlo runs) of the correct classification rate, sensitivity, specificity and area under the curve were equal to 96.42% ± 2.04, 98.69% ± 1.38, 94.16% ± 3.52 and 0.982 ± 0.017, respectively.

Received 13th August 2015,
Accepted 21st December 2015

DOI: 10.1039/c5an01656h

www.rsc.org/analyst

1. Introduction

Models that are constructed using a large number of explanatory variables are prone to the overfitting issue. Namely, they tend to provide very optimistic predictions for the samples used to construct a model, but perform poorly for future samples, the so-called test set. This phenomenon has been well-documented in the literature (see *e.g.* ref. 1) and is often observed when data consisting of instrumental signals are modeled. To deal with the overfitting issue two strategies are usually applied: (i) a careful selection of model complexity and (ii) the selection of relevant variables that support the construction of an adequate model. Both require appropriate validation.

In order to validate the performance of a model it is necessary to gain insight into its predictive abilities.[2] This can be done using different cross-validation/re-sampling methods (*e.g.* leave-one-out, leave-n-out cross-validation, bootstrap, jackknifing, Monte Carlo, *etc.*) and/or an independent set of samples. These model validation concepts are often referred to as internal and external validation. The cross-validation approaches are typically used to estimate the optimal number of components that should be used for the construction of a model. As claimed by Esbensen and Geladi,[2] the only appropri-

[a]*Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland. E-mail: mdaszyk@us.edu.pl; Fax: +48 32 259 99 78; Tel: +48 32 359 1568*
[b]*Scientific Institute of Public Division of Food, Medicines and Consumer Safety, Section Medicinal Products Health (WIV-ISP), Rue Juliette Wytsmanstraat 14, Brussels, 1050, Belgium*
[c]*Research group NatuRA (Natural products and Food – Research and Analysis), Department of Pharmaceutical Sciences, University of Antwerp, Universiteitsplein 1, B-2610 Wilrijk, Belgium*

ate model validation is based on an independent test set. The test set is fully independent – it is never used at any stage of the construction of the model. It can be designed using subset selection approaches, for instance, uniform subset selection methods like Kennard and Stone algorithm and Duplex, clustering methods, random selection, D-optimality criterion, *etc.*[3] The general principles of multivariate model validation can be followed regardless of the intended purpose of a model – calibration, discrimination and/or classification.

In this study, we focused on the validation of the discriminant partial least squares models (PLS-DA) used in the field of drug control to discriminate between authentic and counterfeit medicines based on their impurity profiles. The major motivation for our research was driven by the fact that in the literature that deals with the detection of counterfeit medicines, the role of model validation is often somewhat underestimated or even neglected. Assuming that authentic samples always have much lower levels of impurities compared to counterfeit samples PLS-DA seemed to be a straightforward choice that would allow the many correlated variables that are found in chromatographic and spectroscopic signals (fingerprints) to be dealt with. We introduced a general validation framework that is based on the Monte Carlo approach. In the course of the Monte Carlo procedure, distributions of selected figures of merit (*e.g.*, sensitivity, selectivity, correct classification rate, *etc.*) are obtained from a discriminant model as a function of its complexity. Therefore, one can easily evaluate the performance of models that have different degrees of complexity based on the selected figures of merit and the uncertainties of their estimates. The proposed model validation procedure was especially designed to fulfill the PLSE-DA assumption of balanced model sets. Moreover, we also assumed that model validation should be performed using balanced test sets in order to obtain error estimates for the groups of authentic and counterfeit samples. We also show that the validation scheme can easily be extended by the selection of the relevant variables that play a key role in differentiating authentic and counterfeit Viagra® samples (including the elimination of uninformative variables,[4] the importance of a variable in the projections,[5] selectivity ratio[6] and significance multivariate correlation[7]).

To illustrate the performance of the validation approach, we focused on the issue of the detection of authentic and counterfeit Viagra® samples based on their chromatographic fingerprints of impurities, which were determined using high performance liquid chromatography. A total of 143 samples of Viagra® medicines were analyzed, including 46 authentic and 97 counterfeit samples. The impurity fingerprints were further modeled using PLS-DA with the goal of developing a reliable diagnostic discriminant model.

2. Theory

2.1 Preprocessing of chromatographic impurity fingerprints

Different preprocessing techniques are usually used to enhance the quality and interpretation of chromatographic impurity fingerprints. In addition to the peaks that originate from the analytes of interest, impurity profiles contain noise and baseline components that can influence the comparative analysis. In the course of a chromatographic analysis of complex mixtures, it is difficult to obtain chromatograms that have baseline separated peaks. Therefore, an over-expressed baseline component is often observed in chromatographic signals. When necessary, the baseline of the signal has to be corrected. A large number of baseline correction methods can be used to perform this task. One of these is the penalized asymmetric least squares (PAsLS) method, which has found numerous applications. A detailed description of the algorithm is provided in ref. 8.

Another preprocessing issue of instrumental signals is related to the possible shifts of the corresponding peaks that can be observed for a collection of signals. The presence of peak shifts in chromatographic fingerprints can be induced, for instance, by fluctuations in the instrumental conditions or changes in the chemical composition of an eluent and/or samples. Correlation optimized warping, COW, is one of the many alignment methods that are frequently used to diminish the negative effect of peak shifts.[9] The alignment of chromatographic fingerprints is achieved by the linear stretching and compression of signal sections in such a way that the overall correlation coefficient between the aligned signal and the target signal is maximized.[9] The target signal is considered to be a template for the alignment and usually reflects the highest correlation coefficient with the remaining signals.[10] Once this is done, it is considered to be representative.

In the COW method, each signal is divided into the same number of sections and all of the sections have the same number of sampling points. In order to match corresponding peaks, their shapes are evaluated. The alignment is controlled by two parameters – the number of sections, N, into which the signals are divided and the slack parameter, s, which influences the flexibility of the alignment. The alignment power of COW can be adjusted to a large extent by modifying the N and s parameters. The major advantage of using COW to adjust peak shifts stems from its ability to preserve the area and shape of a peak.

2.2 Exploratory analysis of impurity fingerprints

Principal component analysis, PCA, is an unsupervised technique that is used to visualize and compress multivariate data.[11] It is usually applied to explore the structure of multivariate data by means of low-dimensional projections that enable groups of samples, local changes of data density and/or objects with unique chemical characteristics compared with the majority of the data to be revealed.[12] In the framework of PCA, a data matrix is represented as the product of the score and loading matrices that contain the so-called principal components, PCs, in columns. It is important to stress that PCs are constructed in order to explain the largest part of the data variability.

Scores are linear combinations of the explanatory variables and are mutually orthogonal, whereas loadings are orthonormal, *i.e.* they are mutually orthogonal and have a unit length.

A large absolute loading value of the original variable indicates the significant importance of that variable in the construction of a given principal component. Scores and loadings are used to visualize the data structure. Projections of selected pairs of scores and loadings provide information about any similarities among the samples and variables, respectively. The distances that are observed among samples (characterized by instrumental signals) in the dataspace described by the scores on selected principal components express their chemical similarity. Loading weight values, which are displayed on the loading plot, express the importance of the variables and the level of their mutual correlation.

2.3 Partial least squares discriminant analysis – model construction and validation

The partial least squares discriminant analysis, PLS-DA, is a variant of the classic partial least squares, which is built to distinguish samples from mutually exclusive groups in a linear manner.[13] For a two-class discriminant problem, a dependent variable, y, which drives the construction of the discriminant rule, defines the belongingness of a sample to a given group. For instance, for authentic *vs.* counterfeit samples, samples are usually labeled either as '−1' and '+1' or '0'and '1'.

In the course of model construction for a two-class problem, a logic rule is built using a few latent variables, which are constructed to maximize the description of data variance and at the same time maximize the covariance between a set of latent variables and a dependent variable. In the context of discrimination using the PLS-DA model, this objective is equivalent to minimizing the within-group scatter while maximizing the distance between the centers of groups.[14]

The future performance of any discriminant or classification model is mostly affected by the representativeness of the samples that are used for its construction. These define its effective domain and influence model complexity (*i.e.* the number of latent PLS-DA variables). It is also important to emphasize that the construction of a PLS-DA model requires a balanced modeling set (*i.e.* one containing the same number of samples from each group). This issue was discussed in detail in ref. 15. It was proven that the PLS-DA decision boundary is shifted towards a larger group, and thus it affects predictions of group labels. It becomes apparent that the samples that are used to construct a discriminant model, in fact, define its effective domain and thus should characterize the data variability expected during the model's maintenance. Usually, this is fulfilled by incorporating the most diverse samples selected from each group into the model set, which is based on a uniform subset selection approach (*e.g.*, the Kennard and Stone algorithm or the Duplex algorithm[3,16]).

The selection of the optimal number of PLS-DA latent variables that are necessary to obtain a model with satisfactory prediction properties can be guided by various cross-validation procedures.[17] The performance of a given model with respect to the recognition of the model and test set samples can be scored by several figures of merit, including the root mean square error for model set samples (RMSE) and the root mean square error for test samples (RMSEP). In general, these two figures of merit measure the level of the overall within-group scatter. However, other figures of merit are usually considered in discrimination problems. Among them the most popular are the correct discrimination/classification rate, CCR (the percentage or proportion of samples with a correctly predicted group label using a model) as well as sensitivity (SE) and specificity (SP). In order to obtain estimates of sensitivity and specificity, the number of true positive samples (TP), true negative samples (TN), false negative samples (FN) and false positive samples (FP) are evaluated. A sample from the group labeled '+1' is called a true positive sample when it is recognized as a sample from that group using a model; otherwise, it is a false negative. When a sample from the group labeled '−1' is recognized as a sample from that group using a model, it is called true negative, otherwise, it is a false positive. Sensitivity and specificity are defined as follows:

$$\mathrm{SE} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \quad (1)$$

$$\mathrm{SP} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \quad (2)$$

In order to illustrate the relationship between the true positive and false positive rates and their influence on model parameters, the so-called ROC plots[18] were introduced. A model's performance is then expressed as the area under the convex curve, AUC. The larger it is the better are the predictive properties of a given model, *i.e.* better discrimination power.

To test the reliability of a discrimination or classification model, it is possible to estimate the uncertainty associated with the estimates of certain figures of merit as a function of model complexity. This can be done in the course of the Monte Carlo[17] or the bootstrap procedure.[19] The Monte Carlo approach, MC, assumes the construction of many subsets by drawing samples in a random manner from the available groups. In this way, sources of variability can be simulated. At each step of the MC procedure, a subset of samples is selected and used to construct models with increasing complexity. Then, its figures of merit are calculated for each model using the remaining samples. Since the MC procedure is repeated many times, the distribution of a selected figure of merit can be constructed, thereby providing the possibility to obtain its uncertainty estimates. In the literature, different approaches for model optimization and validation are described, including the MC cross-validation, cross model validation, *etc.*[17,20]

In this study, we adopted the idea of the MC validation specifically for a two-class discriminant problem that is solved using PLS-DA with an assumption of the balanced representation of two groups of samples (and also extended with the selection of relevant variables). A general scheme of the proposed MC validation procedure is presented in Fig. 1.

At the first step, a balanced set of samples is selected at random from the available set of samples (the so-called initial model set). The remaining samples form an unbalanced set that is put aside as external test set. In the course of the MC procedure, the assumed number of samples is drawn randomly (without any replacement) from the initial model set

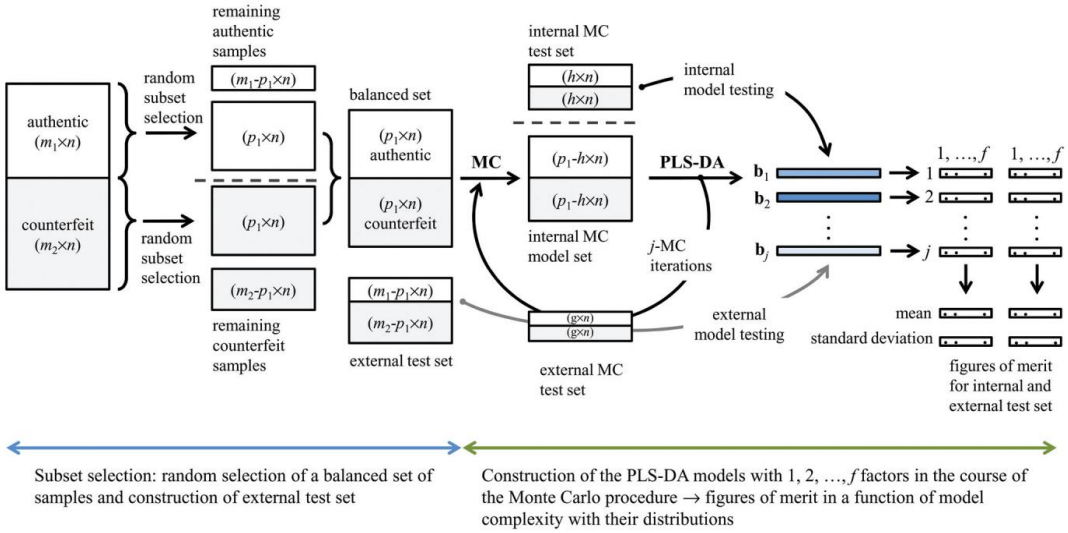


Fig. 1 A general validation scheme for partial least squares discriminant analysis based on the Monte Carlo procedure.

and the pool of external test set samples. PLS-DA models with increasing complexity, 1, 2,..., f, are constructed for a given model subset. Each model is characterized by the selected figures of merit that are obtained for the balanced model set, internal test set and external test set. After j MC runs, the distribution for a given figure of merit and a fixed model complexity are obtained. The final results are reported as the average value, which is extended with the corresponding standard deviation as a function of model complexity. It should be emphasized that the external set of samples is independent with respect to all phases of modeling and variable selection, *i.e.* these samples are never used at the stage of model optimization or during variable selection – they only serve to test model performance.

Such a validation procedure allows the direct estimation of model complexity and straightforward validation using a completely external test set of samples. In order to enable a fair comparison and honest error estimates for two groups of samples, all of the subsets of the samples that are drawn in the course of the Monte Carlo approach are balanced. At the same time, model performance is revealed, which includes information about the uncertainty that is associated with the estimated figures of merit.

2.4 Selection of relevant variables for PLS-DA

Over the last few years, different variable selection approaches have been proposed in order to limit the risk of model overfitting and to enhance interpretation. Some of them, *e.g.*, variable importance in projection (VIP)[5] and significance multivariate correlation (SMC),[7] are specifically designed to support variable selection for partial least squares regression and PLS-DA, whereas uninformative variable elimination and the selectivity ratio are not limited to PLS only.[4,21,22]

2.4.1 Uninformative variable elimination-partial least squares discriminant analysis. Uninformative variable elimination partial least squares discriminant analysis, UVE-PLS-DA, is designed to eliminate variables that do not support the modeling of the dependent variable y. In practice, these variables contain information content that is comparable to the random variables.[4] To distinguish between informative and uninformative variables, the stabilities of the regression coefficients for the original variables and random variables (obtained in the course of the jack-knifing procedure) are compared. Therefore, the UVE-PLS-DA model is built for augmented data that contain the experimental data matrix $\boldsymbol{X}$ of dimensions $m \times n$ (m samples and n variables) and matrix $\boldsymbol{N}$ with random variables that are normally distributed with low magnitudes ($m \times n^*$). Artificial variables do not influence the construction of the UVE-PLS-DA model.

The stability of a regression coefficient is defined as the ratio between the mean value of the regression coefficients for a given variable obtained from the jack-knifing approach and their standard deviation. It is intuitive that informative variables must have absolute stabilities that are larger than the absolute values of the stabilities observed for random variables. Therefore, the threshold value can be selected as, for instance, the maximal value of the absolute stabilities that are found for random variables or a certain percentile value, *e.g.*, 0.99.

Uninformative variables are then discarded and the final PLS-DA model is built based on the remaining variables.

2.4.2 Variable importance in projection. Variable importance in projection, VIP, is a variable selection method that helps to filter out variables in PLS-DA that are irrelevant for a

given discriminant or calibration problem.[23] The importance of variables is scored by the VIP value, which is determined as described in ref. 24. The influence of variables is classified by means of the VIP score as follows: VIP > 1.0 (highly influential), 0.8 < VIP < 1.0 (moderately influential) and VIP < 0.8 (less influential). It is recommended that the VIP criterion be used in a recursive manner, *i.e.* to perform the elimination of variables until no further model improvement is observed.

2.4.3 Selectivity ratio. Another filter variable selection method that can be used to determine the importance of variables in PLS is the so-called selectivity ratio, SR. The selectivity ratio is calculated for each variable as the ratio between the variance explained by the PLS model and the variance of the model residuals.[6,21] A high SR value means that the variable has a strong ability to discriminate between the analyzed groups of samples. The threshold above which the variables are considered as important is arbitrarily chosen by the user.

2.4.4 Significance multivariate correlation method. The significance multivariate correlation method, SMC, assists in assessing the significance of variables in PLS regression or PLS-based discriminant analysis.[7] SMC belongs to the category of filter variable selection methods. Its main aim is to estimate the sources of the relevant variability for each variable based on model regression coefficients (*i.e.* explained variance and residual variance). The importance of variables is described by the SMC parameter, which is calculated for each variable taking into account the predicted response and the regression coefficients that are obtained for a given model. Variables with relatively high SMC values are better correlated with the response variable y and thus they can be regarded as relevant for a given regression or discrimination problem. In order to identify the relevant variables, the threshold value for the SMC values is defined based on the F-test with an assumed significance level of α and 1 and $m - 2$ degrees of freedom. Similar to VIP, SMC is designed to assist in the assessment of the relevance of a variable in the construction of models that use the PLS approach.

2.4.5 Variable selection and the Monte Carlo procedure. All of the variable selection approaches discussed above can be used during the MC procedure to obtain information about the relevance of variables and the frequency of their selection based on multiple model sets. A set of retained variables is stored at each step of the MC procedure. The final set of variables is determined for the model that has optimal complexity assuming a certain frequency of their selection, *e.g.*, variables found in 95% of all of the MC runs (see Fig. 2). Then, the final model that contains only relevant variables is constructed and validated using the general procedure shown in Fig. 1.

3. Experimental

A collection of 46 authentic and 97 counterfeit Viagra® samples was analyzed using a high performance liquid chromatography system (Waters 2695 Separations Module, Milford, USA) with a photo-diode array detector (Waters 2998 Photodiode Array Detector, Milford, USA). The following sample preparation was applied – 30 mg of the sample was dissolved in 10 mL of ethanol/water (50/50% v/v) for 15 minutes using an ultrasonic treatment. Afterwards, the samples were centrifuged at a speed of 2000 rpm for 10 minutes. The supernatant was used for chromatographic analysis. All of the steps of sample preparation were performed at room temperature.

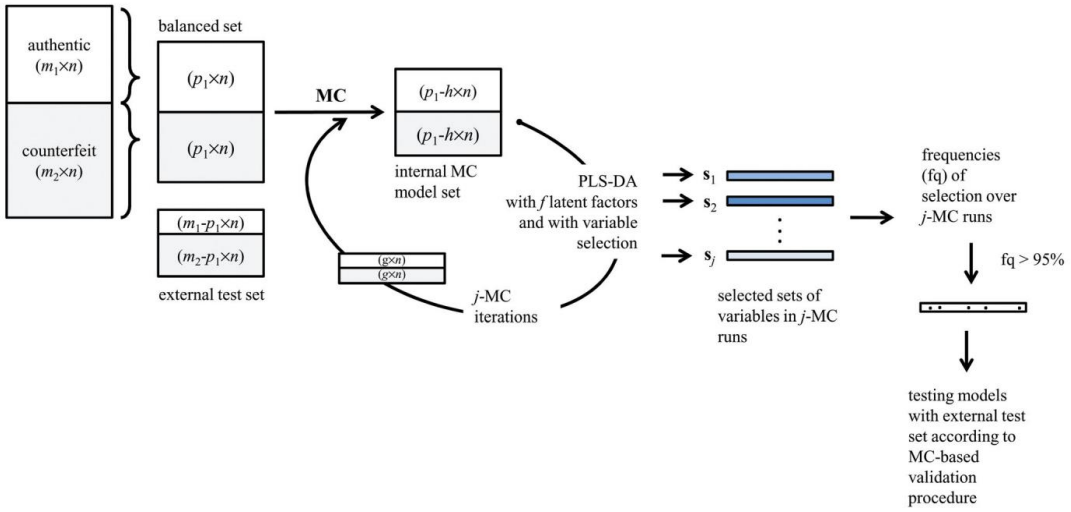


Fig. 2 A general scheme of variable selection embedded into partial least squares discriminant analysis model construction based on the Monte Carlo procedure for the evaluation of variable selection frequency (fq).

Five μl of the sample was injected into the HPLC system. The autosampler temperature was 15 °C and the column temperature was set to 30 °C. Separation was carried out using a C18 column (Alltima, 250 mm × 3 mm; 5 μm particle size; Grace, Columbia, USA) and a binary mobile phase that was composed of an ammonium formate buffer (0.020 M, pH = 3) and methanol. Their proportions were controlled *via* the following gradient program – for two minutes; the mobile phase consisted of 90% of the buffer and 10% methanol. Then, the proportion of the buffer and methanol was linearly altered to 50% for the next five minutes and kept at this level for the next seven minutes. Afterwards, the proportion of the buffer and methanol reached 10% and 90%, respectively, for six minutes and this mobile phase composition remained constant for five minutes. During the last five minutes of separation, the initial mobile phase composition was reached and the total separation run was completed after 30 minutes. A constant mobile phase flow rate of 0.5 ml min^{-1} was applied during the analysis. Spectra were recorded between 210 and 400 nm with a 1.9 nm step for each portion of the eluent.

All data were acquired using the Empower software version 3 (Waters, Milford, USA). The final set that was used for the construction of diagnostic models consisted of chromatograms recorded at the optimal detection wavelength (254 nm). At this wavelength excipients such as like lactose, croscarmellose, *etc.* do not absorb, and therefore they are not detected.

3.1 Hardware and software description

Calculations were performed using a HP ProBook 6560b personal computer with processor Intel(R) Core(TM) i5-2520M CPU 2.50 GHz and 16 GB RAM, Operating System: Microsoft Windows 7 Version 6.1 (Build 7601: Service Pack 1). All algorithms discussed in this manuscript were developed in-house in the MATLAB computing environment (MATLAB Version: 8.1.0.604 (R2013a). The MATLAB routine for validation of PLS-DA using the MC framework and related algorithms are available from the corresponding author upon request.

4. Results and discussion

4.1 Data preprocessing

Since all of the chromatographic fingerprints contained the same number of sampling points (13 620) and were registered within the same range of elution times, they did not require resampling. In order to eliminate any differences in the baseline of the chromatographic fingerprints, the PAsLS method was used. The choice of the λ parameter was optimized. For most of the chromatographic fingerprints, $\lambda = 10^5$ was found to be the optimal value and the second degree of differences was considered.

The next step of preprocessing consisted of the alignment of the signals due to the presence of peak shifts. The misalignment issue was easily detected when the position of the active substance peak was analyzed across the collection of chromatograms. At the step of signal preprocessing, the peak of the active substance served as the marker peak, which additionally helped to verify the alignment. The correlation optimized warping method was used to correct the position of corresponding peaks. The target signal for the alignment was selected as the one that resembled the best average correlation with all of the remaining signals.[10] In order to achieve a satisfactory alignment using COW (in terms of the improvement of the correlation coefficient that was evaluated before and after alignment), different values of the input parameters were tested. In most cases, the values of *ca.* 28 sections ($N = 500$) and $s = 3$ enabled a relatively high alignment flexibility.

Afterwards, alignment tuning in two problematic signal sections was carried out (between 10.10 min and 10.43 min and between 20.10 min and 20.44 min) using the optimal values of the slack parameter that were equal to 3 and 6, respectively. For the first elution time, window N was equal to 20 and for the second 50.

The initial values of the correlation coefficients, which were calculated between the chromatographic fingerprints and the target signals, were in the range of 0.0134 to 0.9988. The initial correlation coefficients did not exceed 0.8 for 48.25% of the chromatograms and they were found to be below 0.9 for 64.34% of the chromatograms. A substantial improvement in peak correspondence was observed after signal alignment. 95.10% of the chromatograms were described by the final correlation coefficient above 0.8 and 90.21% of the chromatograms had a final correlation coefficient above 0.9. In order to visualize the alignment effect, histograms of the correlation coefficients that described individual chromatographic fingerprints before and after alignment are shown in Fig. 3.

Only impurity profiles were considered in this study and therefore after the preprocessing step the peak of the active substance was removed from all signals (*i.e.* the region between 20.22 min and 20.85 min). Afterward, the chromatographic impurity fingerprints of the authentic and counterfeit Viagra® samples were further analyzed using the PCA approach and discriminated using PLS-DA.

4.2 Exploratory analysis of chromatographic impurity fingerprints

Potential differences between the authentic Viagra® samples and their counterfeit variants were explored using the PCA score projections that visualized the similarities among their impurity profiles. In the score projections in Fig. 4, authentic and counterfeit samples are marked as '+' and 'O', respectively. More than 87.86% of the total data variance is explained by the first three principal components. Projection of samples onto a space that was defined by the first two principal components (Fig. 4a) revealed six unique (outlying) samples with very different chemical characteristics in comparison to the remaining samples. These belonged to the group of counterfeit samples. Due to the outlying character of these samples, they were excluded from the data set to eliminate their negative influence on the construction of the discriminant model.

In general, the analysis of score plots led to the conclusion that the group of counterfeit samples was much more inhomo-

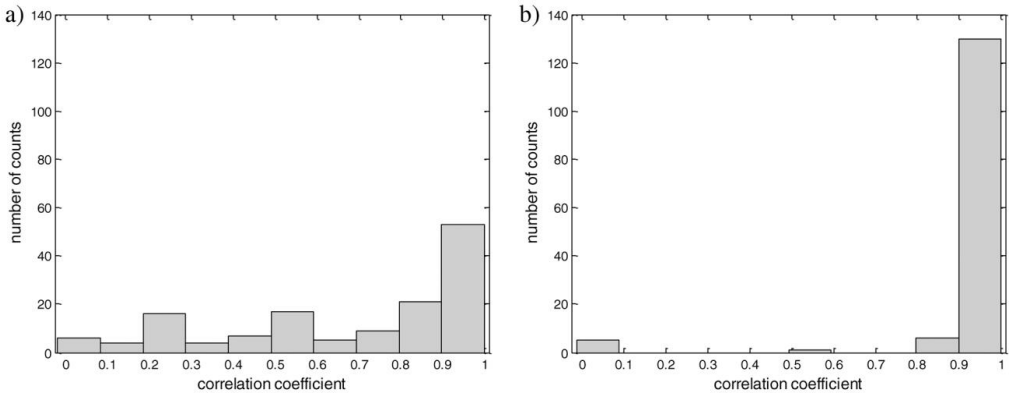


Fig. 3 Histograms of the correlation coefficients that were calculated between a target signal and all of the chromatographic fingerprints: (a) before and (b) after alignment using the correlation optimized method.

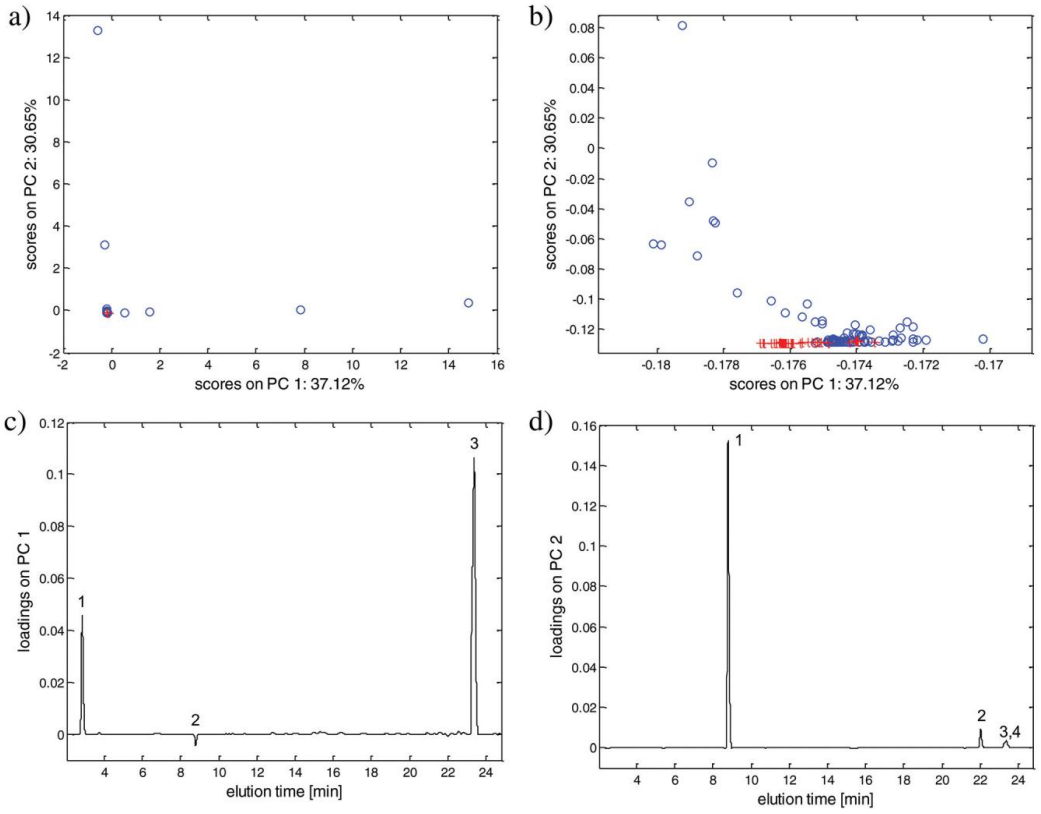


Fig. 4 Score plots of the first two principal components of impurity chromatographic fingerprints: (a) 46 authentic '+' and 97 counterfeit 'O' Viagra® samples, (b) enlarged region of the PC 1–PC 2 score plot, (c) loadings on PC 1 with the three indicated elution regions at (1) 2.855, (2) 8.80 and (3) 23.365 min and (d) loadings on PC 2 with four elution regions indicated at (1) 8.80, (2) 22.02, (3) 23.26 and (4) 23.37 min, where the most influential impurities elute.

geneous and scattered in comparison with the authentic Viagra® samples. This observation is not surprising since the production of counterfeit medicines has little to do with good manufacturing practice and maintaining reasonable levels of production quality. A larger scatter of counterfeit samples, especially along the PC 2 axis, confirmed the hypothesis that illegal counterfeiting practice is indeed a source of the additional variability that is manifested at the chemical level and can be readily explained by the increasing number and/or levels of impurities (see Fig. 4b). In addition, score projections indicated a separation tendency between the authentic and counterfeit samples. Apparently, the largest differences that were observed between authentic and counterfeit samples were due to the presence of the impurities that eluted at *ca.* 23.365 min and 2.855 min (see Fig. 4c). A larger scatter of counterfeit samples along PC 2 was mostly caused by a higher content of impurities that eluted at *ca.* 8.8 min (see Fig. 4d).

4.3 Discrimination of authentic Viagra® samples and their counterfeit variants

In order to build a logic rule that could effectively support the discrimination of the authentic Viagra® samples and their counterfeit variants based on their chromatographic impurity profiles, the PLS-DA method was used. The selection of a linear discriminant method appeared to be a straightforward choice for our pilot study.

Regardless of the type of model that is considered, it is meant to support the decision-making process over a longer period of time. Its maintenance has very much to do with the samples used for its construction. In authenticity studies of different medicines, it is impossible to create or design a model set of samples since they will never reflect the potential variability of counterfeit variants. Therefore, local markets are sampled with the hope that the material that is acquired will describe the expected variability of counterfeit samples as well as possible. The construction of a model requires a representative set of samples that contain representative sets of authentic and counterfeit samples. Generally, the variability of authentic samples will be relatively small. Following this reasoning, in most of the published studies only a few authentic samples have been considered. On the other hand, the group of counterfeit samples is often considerably larger and this may raise two issues. The imbalanced proportions of samples will lead to difficulties in testing the constructed models since the independent test set will contain significantly fewer authentic samples than counterfeit samples. Using imbalanced groups of samples may, depending on the discriminant approach that is applied, influence the construction of the discriminant hyperplane. In support vector machines or *k*-nearest neighbor techniques, the construction of a separating hyperplane effectively involves only the samples that are located at the borders or very close to the borders of the groups. However, in LDA or PLS-DA, the situation is very different. All of the samples are required in order to determine the optimal location of the discriminant hyperplane and thus their proportions and group variances play a fundamental role.

In this study, we used a balanced model, test and external test sets of samples. The following number of samples was considered in the Monte Carlo scheme that is presented in Fig. 1: $m_1 = 46$, $m_2 = 97$, $p_1 = 35$, $h = 10$ and $g = 8$. All of the samples were selected randomly without any replacement. Such a construction of different sets offers the possibility to (i) design them with respect to the number of samples and (ii) simulate the different sources of variability. The MC procedure was repeated 1000 times. In a single step, models with different number of factors were used in order to obtain estimates of the selected figures of merit (the correct classification rate, sensitivity, specificity and AUC) for a given configuration of the internal model set, test set and external test set.

After the MC procedure, a distribution of the selected figures of merit is available for a given model complexity, and thus their uncertainties can be estimated. Estimates of the correct classification rate are presented in Fig. 5, as the average value of correct classification rates extended with their standard deviation (vertical bars) from 1000 MC runs as a function of model complexity.

An analysis of the CCR values that were obtained in the course of the Monte Carlo procedure for the internal test set suggested that the optimal PLS-DA model should contain 5 PLS factors, thus leading to 89.37% ± 1.48 of the correct classification rate for the model set, 90.60% ± 3.97 for the internal test set, and 88.03% ± 2.64 for the external test set. Additionally, the relatively high values of sensitivity, specificity and AUC that were obtained for the external test set confirmed that the discriminant problem that was studied could be solved with a simple linear PLS-DA model (see Fig. 4b–d). A detailed presentation of all of the figures of merit that were considered is provided in Table 1.

4.4 Variable selection

As was confirmed by the relatively high values of the figures of merit that are presented in Table 1, the PLS-DA model can differentiate authentic Viagra® samples and their counterfeit variants with great success. On the other hand, the large number of variables that were used for the construction of the model compared to the number of available samples increased the risk of model overfitting and complicated the identification of regions where the impurities, which are relevant for discrimination, eluted. Therefore, variable selection is usually recommended in order to limit model over optimism.[25]

Several variable selection methods that can be easily embedded into the PLS-DA framework were considered to eliminate uninformative chromatographic features. These included uninformative variable elimination, variable importance in projection, the selectivity ratio and significance multivariate correlation. The final PLS-DA models were constructed using subsets of the relevant variables. Their performance was evaluated using the MC procedure, which was repeated 1000 times and was tested with the external test set.

In the UVE procedure, only variables with absolute stabilities above the 99% percentile of the absolute stabilities were retained at each Monte Carlo step. The final set of variables

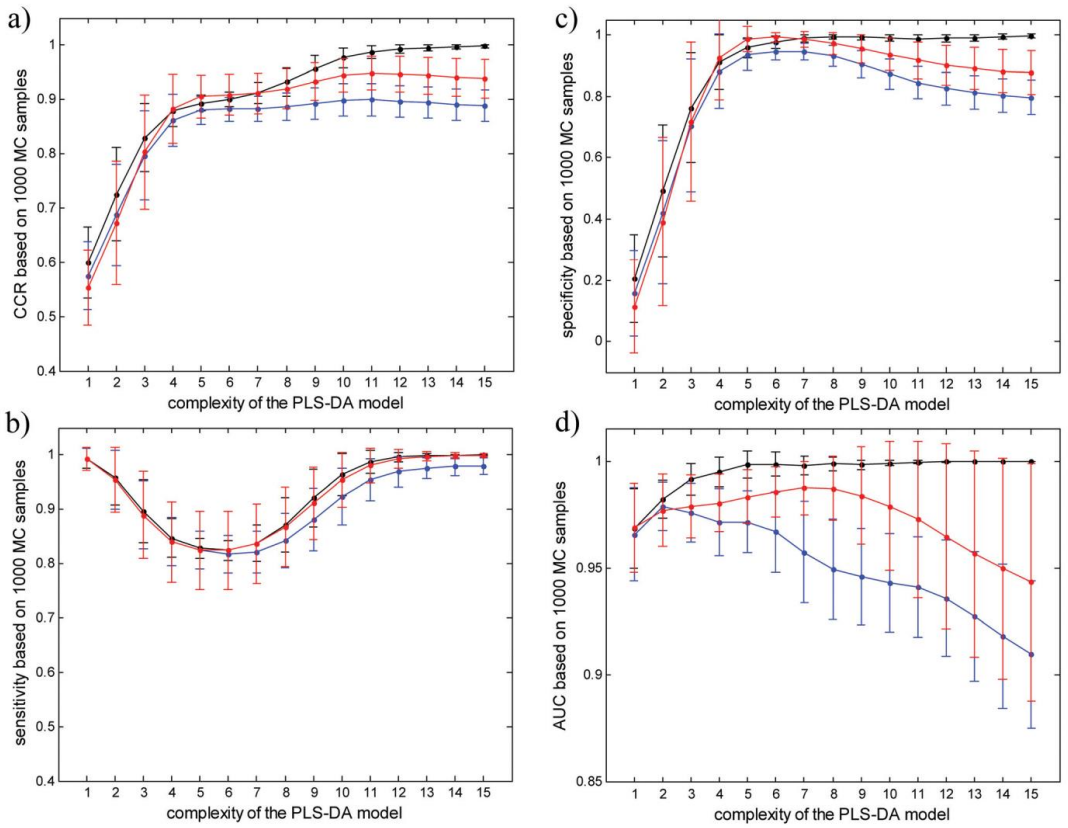


Fig. 5 (a) Correct classification rates (CCR), (b) sensitivity (SE), (c) specificity (SP) and (d) area under curve (AUC) that were obtained as a function of the PLS-DA model complexity extended with uncertainty estimates expressed as standard deviation (vertical lines) obtained in the course of the Monte Carlo procedure (1000 runs) for the internal model set (black line), the internal test set (red line) and the external test set (blue line).

contained variables that were identified as relevant in all of the MC runs. For VIP, only the variables selected in all of the runs were considered and the VIP procedure was applied recursively three times. Using the selectivity ratio, the threshold value was set to 0.9 and the final set of variables contained the variables selected in 95% of all MC runs (higher threshold values or higher selection frequencies resulted in an empty set of the selected variables). In SMC, the final set of variables contained variables characterized by 100% of selection frequency.

Analysis of the results that are presented in Table 1 allowed it to be concluded that, in general, the variable selection procedures that were applied decreased the complexity of all of the PLS-DA models. For a subset of variables selected using either SR or SMC, the final models contained one factor less compared to the initial model. The subsets of the variables selected with UVE and VIP resulted in a larger reduction of model complexity – from five to two factors. It is important to emphasize that the reduction of complexity for the studied data has no negative effect on the prediction properties of the models (see Table 1). Of all of the variable selection methods that were applied to the data, SR had the most restrictive performance in terms of the number of discarded variables. Only 21 of the 13 291 variables were detected as relevant to the differentiation of authentic and counterfeit drug samples. The largest set of variables, which contained 3641 variables, was retained using SMC. All of the constructed models, with and without a variable selection step, had average correct classification rates of above 88% for the external test set with an uncertainty estimate below 2.86%. The best PLS-DA model, in terms of CCR, SE, SP and AUC estimates for the external test set, was constructed for the subset of variables selected using VIP (see Table 1). In general, one can conclude that regardless of the PLS-DA model with the VIP-based variable selection, all of the models had a tendency to describe the counterfeit samples better. This is supported by the larger specificity

Table 1 Different figures of merit (CCR – correct classification rate, SE – sensitivity, SP – specificity and AUC – area under curve) that were obtained from the classic PLS-DA model and PLS-DA extended with four different variable selection schemes (UVE – uninformative variable elimination, VIP – variable importance in projection, SR – selectivity ratio and SMC – significance multivariate correlation) embedded into the Monte Carlo-based validation. Values of different figures of merit were obtained for model set samples, internal test set and independent test set samples randomly drawn 1000 times in the course of the Monte Carlo procedure. Each figure of merit is reported as the average value over 1000 runs and accompanied by the estimate of uncertainty (standard deviation). Symbols f and k denote the number of latent PLS-DA factors and the number of considered (or selected using a variable selection technique) explanatory variables, respectively

			Monte Carlo model set				Monte Carlo internal test set				Monte Carlo external test set			
Model	f	k	CCR [%]	SE [%]	SP [%]	AUC	CCR [%]	SE [%]	SP [%]	AUC	CCR [%]	SE [%]	SP [%]	AUC
PLS-DA	5	13 291	89.37 ± 1.48	82.82 ± 1.88	95.92 ± 3.25	0.999 ± 0.006	90.60 ± 3.97	82.44 ± 7.31	98.75 ± 4.06	0.984 ± 0.013	88.03 ± 2.64	82.48 ± 3.48	93.58 ± 5.03	0.972 ± 0.015
UVE	2	674	89.11 ± 1.44	82.47 ± 1.44	95.76 ± 2.62	0.960 ± 0.013	90.74 ± 3.70	81.64 ± 7.39	99.83 ± 0.56	0.891 ± 0.054	88.36 ± 2.19	82.45 ± 3.28	94.28 ± 3.40	0.940 ± 0.026
VIP	2	83	97.84 ± 1.39	99.10 ± 1.04	99.57 ± 2.26	0.999 ± 0.000	93.34 ± 3.32	100.00 ± 0.00	86.68 ± 6.65	0.956 ± 0.047	96.42 ± 2.04	98.69 ± 1.38	94.16 ± 3.52	0.982 ± 0.017
SR	4	21	91.32 ± 0.69	82.63 ± 1.37	100.00 ± 0.00	0.955 ± 0.008	90.71 ± 3.73	81.64 ± 7.39	99.79 ± 0.65	0.889 ± 0.046	89.72 ± 1.90	81.26 ± 3.40	98.19 ± 1.60	0.936 ± 0.022
SMC	4	3641	94.22 ± 1.95	91.47 ± 3.76	96.97 ± 2.79	0.986 ± 0.009	94.41 ± 3.11	90.30 ± 6.39	98.52 ± 2.58	0.999 ± 0.003	91.38 ± 2.86	88.71 ± 5.44	94.05 ± 4.35	0.962 ± 0.016

values compared with the corresponding selectivity values. On the other hand, the PLS-DA model with the VIP-based variable selection offered the best performance for the authentic Viagra® samples (SE = 98.69% ± 1.38).

4. Conclusions

In this paper a general framework for the validation of PLS-DA models, which were built to discriminate between authentic and counterfeit samples, was proposed. It takes into account balanced data representations based on the MC approach. Such an approach enables a simulation of variability due to the random selection of samples and at the same time the observation of model performance up to a maximal considered complexity. The major advantage stems from the possibility of obtaining distributions of figures of merit as a function of model complexity. It is also possible to easily extend the proposed framework using a variable selection procedure, *e.g.*, VIP, SR, UVE or SMC. In general, such a strategy assists in reducing the over optimism of the PLS-DA models that are constructed and enhances their interpretation. The selected chromatographic variables (regions along the elution time axis) can help in the further identification of potential chemical markers of the counterfeiting processes that are studied using a complementary analytical technique. This strategy can also be used in other studies related to authenticity confirmation, which are designed to uncover differences between authentic and non-authentic samples at the chemical level, for instance to detect illegal fuel discoloration.[26] In general, all of the discussed PLS-DA models (with and without a variable selection scheme) offered a relatively high prediction performance. The best diagnostic model was based on PLS-DA constructed for a subset of variables selected using the variable importance in the projection approach. The average estimates with corresponding standard deviations for the independent test set (based on 1000 Monte Carlo runs) for the correct classification rate, sensitivity, specificity and area under curve were equal to 96.42% ± 2.04, 98.69% ± 1.38, 94.16% ± 3.52 and 0.982 ± 0.017, respectively.

In general, the proposed validation workflow could also be used in many other discrimination and classification tasks, for instance, food adulteration or food authenticity studies (and not only) solved with techniques other than PLS-DA discriminant models or class modelling techniques.

Acknowledgements

MD wishes to acknowledge the support of the National Science Centre, Poland (research grant no. 2014/13/B/ST4/05007).

References

1 N. M. Faber and R. Rajkó, *Anal. Chim. Acta*, 2007, **595**, 98–106.

2 K. H. Esbensen and P. Geladi, *J. Chemom.*, 2010, **24**, 168–187.
3 M. Daszykowski, B. Walczak and D. L. Massart, *Anal. Chim. Acta*, 2002, **468**, 91–103.
4 V. Centner, D.-L. Massart, O. E. De Noord, S. De Jong, B. M. Vandeginste and C. Sterna, *Anal. Chem.*, 1996, **68**, 3851–3858.
5 O. M. Kvalheim, R. Arneberg, O. Bleie, T. Rajalahti, A. K. Smilde and J. A. Westerhuis, *J. Chemom.*, 2014, **28**, 615–622.
6 O. M. Kvalheim, *J. Chemom.*, 2010, **24**, 496–504.
7 T. N. Tran, N. L. Afanador, L. M. C. Buydens and L. Blanchet, *Chemom. Intell. Lab. Syst.*, 2014, **138**, 153–160.
8 P. H. C. Eilers, *Anal. Chem.*, 2003, **75**, 3631–3636.
9 N.-P. V. Nielsen, J. M. Carstensen and J. Smedsgaard, *J. Chromatogr., A*, 1998, **805**, 17–35.
10 M. Daszykowski and B. Walczak, *J. Chromatogr., A*, 2007, **1176**, 1–11.
11 S. Wold, K. Esbensen and P. Geladi, *Chemom. Intell. Lab. Syst.*, 1987, **2**, 37–52.
12 M. Daszykowski, B. Walczak and D. L. Massart, *Chemom. Intell. Lab. Syst.*, 2003, **65**, 97–112.
13 M. Barker and W. Rayens, *J. Chemom.*, 2003, **17**, 166–173.
14 E. K. Kemsley, *Chemom. Intell. Lab. Syst.*, 1996, **33**, 47–61.
15 R. G. Brereton and G. R. Lloyd, *J. Chemom.*, 2014, **28**, 213–225.
16 R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137–148.
17 Q.-S. Xu, Y.-Z. Liang and Y.-P. Du, *J. Chemom.*, 2004, **18**, 112–120.
18 T. Fawcett, *Pattern Recognit. Lett.*, 2006, **27**, 861–874.
19 R. Wehrens, H. Putter and L. M. C. Buydens, *Chemom. Intell. Lab. Syst.*, 2000, **54**, 35–52.
20 E. Szymańska, E. Saccenti, A. K. Smilde and J. A. Westerhuis, *Metabolomics*, 2012, **8**, 3–16.
21 T. Rajalahti, R. Arneberg, A. C. Kroksveen, M. Berle, K.-M. Myhr and O. M. Kvalheim, *Anal. Chem.*, 2009, **81**, 2581–2590.
22 M. Farrés, S. Platikanov, S. Tsakovski and R. Tauler, *J. Chemom.*, 2015, **29**, 528–536.
23 R. Gosselin, D. Rodrigue and C. Duchesne, *Chemom. Intell. Lab. Syst.*, 2010, **100**, 12–21.
24 S. Favilla, C. Durante, M. L. Vigni and M. Cocchi, *Chemom. Intell. Lab. Syst.*, 2013, **129**, 76–86.
25 C. M. Andersen and R. Bro, *J. Chemom.*, 2010, **24**, 728–737.
26 B. Krakowska, I. Stanimirova, J. Orzel, M. Daszykowski, I. Grabowski, G. Zaleszczyk and M. Sznajder, *Anal. Bioanal. Chem.*, 2015, 1–12.

Brussels, 21/06/2016

Deborah Custers, PhD
Scientific Institute of Public Division of Food
Medicines and Consumer Safety
Section Medicinal Products Health (WIV-ISP)
Juliette Wytsmanstraat 14
Brussels
Belgium

To whom it may concern

Hereby,

I declare that my overall contribution to the article entitled "The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles" by B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, published in Analyst, 141 (2016) 1060-1070 my contribution mostly concerned:

- developing a chromatographic procedure for profiling impurities of Viagra® samples and their counterfeit variants,
- the preparation of samples and developing chromatographic profiles,
- data transfer and data organization,
- providing comments on the manuscript content prior to its submission,
- participating in preparing answers to reviewers' and assistance in preparing a revised version of the manuscript.

Deborah Custers

21-06-2016

WETENSCHAPPELIJK INSTITUUT
VOLKSGEZONDHEID
INSTITUT SCIENTIFIQUE
DE SANTÉ PUBLIQUE

Section of Medicines

datum : 21/06/2016
uw ref. :
onze ref.

contact : Eric Deconinck
tel. : + 32 2 642 51 70
fax : + 32 2 642 53 27
e-mail : Eric Deconinck@wiv-isp.be

To whom it may concern

Hereby,

I declare that my overall contribution to the article entitled "The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles" by B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, published in Analyst, 141 (2016) 1060-1070 mostly concerned:

- developing chromatographic procedure for profiling impurities of Viagra® samples and their counterfeit variants,
- providing chromatographic data,
- interpretation of obtained result from a pharmaceutical point of view,
- providing comments on the manuscript content prior to its submission,
- participating in preparing answers to reviewers' and assistance in preparing a revised version of the manuscript.

Pharm. Eric Deconinck, PhD
Head of section
Section of medicines
Operational Direction Food
Medicines and Consumer Safety

Juliette Wytsmanstraat 14
1050 Brussel | België
T + 32 2 642 51 11 | F + 32 2 642 50 01
info@iph.fgov.be | www.iph.fgov.be

dr hab. Michał Daszykowski, prof. UŚ
Instytut Chemii
Uniwersytet Śląski
ul. Szkolna 9
40-006 Katowice

Katowice 29.06.2016

Oświadczam, że w artykule pt. „The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity" opublikowanym w czasopiśmie Analyst, 141 (2016) 1060 mój udział polegał na:

- współtworzeniu hipotezy badawczej,
- pomocy w opracowaniu i przeprowadzaniu modyfikacji proponowanej procedury walidacji modeli PLS-DA,
- pomocy w interpretacji wyników uzyskanych przez doktorantkę,
- opiece i merytorycznym nadzorze procesu przygotowania manuskryptu,
- pomocy w przeprowadzeniu procedury redakcyjnej i przygotowaniu odpowiedzi na recenzje,
- dokonaniu ostatecznej korekty artykułu.

Michał Daszykowski

Załącznik 3

Publikacja:	Expert system for monitoring the tributyltin content in inland water samples
Autorzy:	M. Daszykowski M. Korzeń B. Krakowska K. Fabiańczyk
Czasopismo:	Chemometrics and Intelligent Laboratory Systems
Wartość współczynnika Impact Factor	2,217

Chemometrics and Intelligent Laboratory Systems 149 (2015) 123–131

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Expert system for monitoring the tributyltin content in inland water samples

M. Daszykowski [a], M. Korzen [b], B. Krakowska [a], K. Fabianczyk [c,*]

[a] Institute of Chemistry, University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland
[b] Faculty of Computer Science and Information Technology, West Pomeranian University of Technology, 49 Zolnierska Street, 71-210 Szczecin, Poland
[c] Polcargo International, Poboznego Street, 70-900 Szczecin, Poland

ARTICLE INFO

Article history:
Received 1 May 2015
Received in revised form 20 October 2015
Accepted 22 October 2015
Available online 30 October 2015

Keywords:
Expert system
Discriminant models
Machine learning
Chemometrics

ABSTRACT

In this study we discuss an attempt to build an expert system that can support decision making by analytical chemists regarding the presence of tributyltin (TBT) in inland Polish water samples in detail. It is possible to conclude with at least a 0.93 probability that a sample is free of TBT using the expert system that was constructed (if a sample is analyzed in accordance with the European norm PN-EN ISO 17353:2006). This idea, which is based on the efficient use of the information that is stored in a chromatographic database, can easily be extended to monitor other priority substances in water samples. Our on-going research, which is focused on octylphenols in water samples, has provided very encouraging results and additionally supports this hypothesis. The proposed framework can also be attractive to other testing laboratories that have a similar scope of expertise and follow the same analytical protocols. Moreover, as a natural consequence of our research further efforts should lead to the development of a ready-to-use product that would offer testing laboratories validated chromatographic libraries along with the expert system(s) with the possibility of upgrading them with respect to an increasing pool of analyzed samples. Such a solution when implemented in a testing laboratory environment may have a wide economic impact on its further functioning and increase throughput efficiency, especially in a case in which monitoring priority substances in water is a major concern.

1. Introduction

Optimizing the costs of demanding analytical procedures while preserving their specificity and reliability is a very challenging task and is a core part of developing intelligent laboratory systems/procedures and management. In general, analytical procedures are time consuming and highly experienced analytical personnel and sophisticated equipment are required.

For decades, different researchers have been seeking appropriate analytical procedures that would enable the identification of substances in very complex analytical matrices, which would then lead to reliable results from an analytical point of view. From this perspective, environmental water samples are a typical example of complex and demanding samples that require separation techniques for their examination. The identification of a large group of chemical constituents in complex mixtures, including environmental water samples, can be done using chromatographic methods and is one of the major challenges in any testing laboratory. Apart from its high level of sophistication, chromatographic analysis is extremely susceptible to external factors, which may result in the shifting of peaks and/or their overlapping and thus complicate the ability to draw conclusions from chromatograms. That is why continuous efforts are undertaken to make chromatographic analysis as reliable and effective as possible. From a practical point of view, two concepts for processing chromatographic signals/data can be identified. The first one relies on the active use of the hardware configurations and options that are available due to the technological advancement of modern chromatographic devices. The second one, which follows a chemometric philosophy, involves the extraction of useful information from complex chromatographic data as a result of the development and implementation of advanced algorithms and their application during different steps of analytical workflow, see e.g., [1–4].

In this study, we contrasted different theoretical approaches with the aim of developing an efficient expert system that is based on machine learning and is proposed to support the detection of the contaminant tributyltin (TBT) in Polish marine and fresh water ecosystems. Its major advantage relies on incorporating the knowledge of experts and a diverse representation of environmental water samples.

Tributyltin, which is a biocide agent, has been extensively used as an ingredient of antifouling paint and is designed to prevent or slow down the growth of organisms on coated with a painted surface. Because it is extremely efficient as a biocide agent, TBT has been used for 40 years mainly in the shipping industry. Initially, it was considered to be environmentally safe, but it was proven that when TBT is released into the

* Corresponding author. Tel.: +48 91 434 33 68; fax: +48 91 488 20 36.
E-mail address: k.fabianczyk@polcargo.pl (K. Fabianczyk).

http://dx.doi.org/10.1016/j.chemolab.2015.10.008
0169-7439/

environment, it exceeds acute and chronic toxic levels. For this reason, different international regulations have been issued to effectively prohibit the further use of TBT and thus to reduce the progression of water contamination with TBT and its degradation products [6]. Unfortunately, toxic effects can still be observed because TBT has a relatively long half-life that depends on its sources, degradation products and their accumulation in sediments, environmental conditions and other factors [7,8]. This is why the level of TBT, including the TBT that is found in different water bodies, is the subject of strict on-going monitoring.

In this study, environmental water samples were collected in the course of the large-scale environmental diagnostic monitoring of inland waters in Poland that was requested by the Chief Inspectorate of Environmental Protection, which was focused i.a. on the detection and quantification of TBT. The sampling campaign took place between 2011 and 2013. TBT and other organotin compounds can be quantified in water samples using the diverse set of available analytical techniques [5]. Among them there are, e.g., the GC–MS method, the headspace solid-phase microextraction-gas chromatography-pulsed flame-photometric detection [6] and the fluorescence technique followed by chemometric modeling using the second-order calibration, which helps in the quantification of TBT at parts-per-trillion levels [7]. In 1403 water samples, the TBT content (quantified as tributyltin cation) was determined using the GC-MS technique according to the European norm PN-EN ISO 17353:2006. The chromatographic fingerprints of water samples that were obtained, despite the rigorous protocols that were applied during chromatographic analysis, reflect all of the major sources of variability. They are rather noisy and contain a substantial baseline component. Moreover, from sample to sample chromatographic peaks are shifted and the overlapping of peaks is frequently observed. On the other hand, real GC-MS fingerprints of complex mixtures pose a real challenge for chemometric and machine learning approaches. As a result of a considerable analytical effort, a relatively large collection of diverse water samples was analyzed. The set of 1403 chromatograms that was obtained offered a unique opportunity to verify their usefulness as a database that would support the construction of an expert system to facilitate in the detection of TBT contaminants in water.

2. Materials and methods

2.1. Sample collection and chromatographic analysis

The sampling plan, sampling frequency, as well as the protocols for sample collection regarding the determination of certain priority substances in the course of the diagnostic monitoring of Polish inland waters in 2011 and 2013 were prepared by the Chief Inspectorate of Environmental Protection in Poland. The following procedure was applied to the analysis of the TBT content in the water samples. Water samples (each 1000 ml) were collected in dark glass flasks. Samples were stored at 4 °C (also during transport). Further sample treatment and analysis was carried out in a specialized laboratory that has up-to-date certificates of accreditation issued by the Polish Center of Accreditation within the scope that includes the analysis of TBT done in accordance with the European norm PN-EN ISO 17353:2006.

Chromatographic fingerprints were registered using a gas chromatographic system (Agilent Technologies 7890A) with a single quadruple mass detector (Agilent Technologies 5975C) with electron ionization. Mixture components (1 μL of a water extract) were resolved using helium as the gas carrier and a DB-5 column (30 m × 250 μm × 0.25 μm). The temperature of the inlet was set to 250 °C and during the separation a temperature gradient was applied (from 60 °C up to 170 °C every minute by 12 °C, then 170 °C up to 280 °C every minute by 20 °C) in the oven. The following ions were monitored (scanning rate at 4.53 cycles/sec): the target ion of tributyltin (TBT) m/z = 291; the qualifier ion tri-n-propyltin (TprT) m/z = 289; the target ion m/z = 249 and the qualifier ion m/z = 247 (temperatures of the transfer line, MS source and MS quadruple were set to 300 °C, 230 °C and 150 °C, respectively and the energy of electrons was set to 69.0 eV).

2.2. Preprocessing of chromatographic fingerprints

A typical workflow for preprocessing chromatographic fingerprints that is carried out prior to multivariate data modeling usually includes the improvement of the signal-to-noise ratio. It consists of baseline removal, noise elimination and the alignment of chromatographic signals [1]. Correction of heteroscedasticity usually is done by transforming data, e.g., logarithm or power transformations [8]. In our study, baseline was corrected using the penalized least squares asymmetric least squares approach, PALS [9]. The alignment of signals was carried out using correlation optimized warping, COW [10]. More details about these preprocessing methods can be found in cited references.

2.3. Discriminant models

The aim of discriminant models is to assign a sample to one of the existing groups of samples based on its characteristics (e.g., a chromatographic fingerprint). In this study, we focused on the discrimination between two groups of water samples with and without TBT. This task essentially corresponds to the issue of identification — the presence or absence of TBT in water samples (confirmed for model set samples by the presence of characteristic mass spectra).

For the pilot discrimination of the groups of samples that were studied, a classic chemometric linear discriminant approach was used — partial least squares–discriminant analysis, PLS-DA. In addition, in order to confirm the presence of TBT in the environmental water samples, the following machine learning methods were used: logistic regression (LR) with the L^1 regularization, linear support vector machines (LK-SVM), ensemble methods including AdaBoost (AB) and random forest (RF), K-nearest neighbors (KNN) and the Parzen classifier (PC).

In the following sections, a brief characteristic of each machine learning technique will be provided.

2.3.1. Partial least squares-discriminant analysis

Partial least squares-discriminant analysis, PLS-DA, is a variant of the classic partial least squares regression model, in which a categorical dependent variable that indicates to which group a sample belongs, is modeled [11]. Any discrimination between the groups of samples is achieved by the construction of a linear separation hyper plane in the space of a few latent variables, which are also called latent or PLS factors. They are mutually orthogonal and maximize the covariance between the set of latent variables and the response variable (in a simple PLS variant with one response, PLS-1). Owing to the construction of orthogonal latent factors, the construction of the PLS model is not hampered by the presence of collinear explanatory variables. In fact, PLS-DA and linear discriminant analysis share the same objective — minimizing within group variance and maximizing between group variance [12].

2.3.2. Logistic regression and family of linear classifiers

PLS-DA is one example from a large group of linear discriminant methods. Others, which are commonly used family of methods within this group, are methods that are based on penalized logistic regression. Similar to PLS-DA, in logistic regression the likelihood function with a penalizing term is used instead of the least squares cost function.

$$Q(\mathbf{X}, \mathbf{y}; \mathbf{b}) = \text{Loglikelihood}(\mathbf{X}, \mathbf{y}) + P(\mathbf{b}) \quad (1)$$

Depending on the type of regularization approach that is selected, a considerably different behavior of a learning machine can be obtained. For instance, the L^2 regularization (or ridge, $P(\mathbf{b}) = ||\mathbf{b}||_2$) leads to a grouping effect of the correlated variables, the L^1 regularization (or

lasso, $P(\mathbf{b}) = ||\mathbf{b}||_1$) results in the automatic selection of variables and the construction of sparse models [13,14]. The so-called elastic net model (weighted sum of the L^1 norm and the L^2 norm of weights as the penalizing term, $P(\mathbf{w}) = \alpha \cdot ||\mathbf{b}||_2 + (1-\alpha) \cdot ||\mathbf{b}||_1$) offers the joint advantage of two features — variable grouping and model sparsity [15]. The α parameter is usually selected using a cross validation procedure. These unique properties of regularization methods can be very useful in the context of modeling chromatographic data with a large number of correlated variables and/or redundant variables.

Support vector machines, SVMs, which were introduced by Vapnik [16], is another popular machine learning technique that is slightly different than the logistic regression. In SVM the separation margin as the loss function is used and the direct minimization of a weight vector via the constrained quadratic programming optimization.

Each linear technique can be transformed into its kernel-based variant when a discriminant problem requires the construction of non-linear boundaries among groups of samples. SVM is a typical example of the kernel-based method, in which linear or non-linear properties can be obtained depending on the type of kernel that is applied [17]. Such a 'kernel trick' can also be helpful in extending the applications of simple linear discriminant techniques in order to deal with non-linear classification problems, e.g., radial basis functions partial least squares [18]. However, when the proportion between the number of variables and samples is large, kernel-based transformations do not improve the accuracy of classifiers substantially.

2.3.3. Ensemble methods

In comparison to classic PLS-DA or logistic regression, there are machine learning algorithms that can offer improved performance by the combined use of a multiple number of "weak" models and the final result is obtained as the weighted sum of outputs from individual classifiers. This approach is referred to as ensemble classifiers. Among the different ensemble-based methods, there are AdaBoost [19] and random forests [20]. They often outperform other methods and are also attractive when the proportion between the number of variables and samples is relatively large.

2.3.4. Instance-based classifiers

Another popular group of machine learning methods is the so-called instance-based classifiers. In this study, the k-nearest neighbor classifier and the Parzen classifier (a variant of kernel-based methods) were used. Instance-based classifiers are constructed using all of the available data. Their performance depends on such parameters as the number of neighbors, k, i.e. the size of neighborhood or the spread of the kernel function (σ). Parameters are typically selected via the cross validation procedure. Instance-based classifiers result in non-linear decision boundaries, which are built locally; however, larger values of k and sigma make the decision boundary smoother.

2.3.5. The variable selection issue

A large number of variables and the high level of correlation among them (observed in chromatographic data) require the use of variable selection or some kind of regularization at an early stage of model construction. Most of panelized models (including the logistic regression and SVMs) can withstand a large number of variables. However, it is known [21] that models that are based on the L^1 regularization have a lower sample complexity than models with the L^2 regularization. The elastic net or lasso regression can be successfully used for the purpose of variable selection. However, it is also possible to embed a variable selection scheme during the construction of PLS and/or ensemble models. In the context of the PLS-DA model, a number of relatively easy to implement filters can be used that help in detecting irrelevant variables. Among them are, for instance variable importance in projection [22], the selectivity ratio [23] and significance multivariate correlation [24]. Some of classifiers, including the K-nearest neighbors or the Parzen classifier, require the selection of relevant variables in the first step in order to achieve a satisfactory performance.

2.4. Validation of discriminant models

Regardless of the type of discriminant model that is considered, its fundamental goal is to provide accurate predictions for new samples. A model's performance can be affected by different factors including the selection of training samples (model set), the selection of model complexity (or other input parameters), a large number of redundant variables, the presence of outlying samples, etc. Therefore, information about its practical value for a future user, which is measured with some meaningful figures of merit (e.g., correct classification rate, sensitivity, specificity, and area under the receiver-operator curve), is crucial.

In our study, the bootstrap approach was used to assess (i) how well the available data support the modeled relationship, (ii) the stability of a model and (iii) the accuracy of its predictions in terms of the uncertainty estimates that were obtained for the model set and the independent test set. The aim of the bootstrap approach is to construct a distribution of certain model parameters based on the available data [25]. It is done by iteratively drawing samples from the available groups in a random manner, which is then followed by the construction of a discriminant model that is characterized by figure(s) of merit. Repeating this process many times helps in obtaining the distribution of a given figure of merit and thus in estimating its uncertainty. The following model parameters are user: the correct classification rate, sensitivity and specificity. CCR is defined as the ration between the number of correctly classified samples using a given model and the total number of samples. Correct classification rate (CCR), sensitivity (SE) and specificity (SP) of a model are calculated based on the number of true positive samples (TP), true negative samples (TN), false positive samples (FP) and false negative samples (FN) as follows:

$$CCR = (TN + TP)/(TN + FP + FN + TP) \quad (2)$$

$$SE = TP/(TP + FN) \quad (3)$$

$$SP = TN/(TN + FP) \quad (4)$$

2.5. Algorithms

Chromatographic signals were aligned using the COW method as implemented by Tomasi et al. in MATLAB, which is available from (http://www.models.kvl.dk/DTW_COW). Logistic regression and support vector machine models were constructed using the *LibLinear* library [26]. Random forests and AdaBoost models were constructed using the *Scikit-learn toolbox* programmed in Python [27]. The remaining procedures and models were constructed using routines that were developed in house.

3. Results and discussion

3.1. Preprocessing of chromatographic fingerprints

The chromatographic fingerprints that were collected revealed the presence of a large number of chemical substances and confirmed the environmental character of the water samples being studied. Each chromatogram contained many chromatographic peaks. Often, their resolution was far from satisfactory in some retention time regions. In addition, despite maintaining strictly controlled chromatographic conditions during the analysis, chromatographic fingerprints contained a substantial baseline component and were noisy. The shape of the baseline component varied from sample to sample. In general, it could be concluded that its intensity increased with the elution time. In regions where the signal had a high intensity, the level of signal noise was

different, which was a strong indication of the presence of heteroscedastic noise. Although the samples that were studied were very complex and often chromatographic profiles resembled large differences due to chemical content, the impression was that peak shifting occurred over time. All of these issues were subjects of concern and required extensive preprocessing prior to the construction of the discriminant model. An exemplary raw chromatographic fingerprint of environmental water sample, which is depicted in Fig. 1, illustrate the high degree of complexity of the signals.

The standard preprocessing of individual chromatographic fingerprints included the enhancement of the signal-to-noise ratio by means of noise and baseline correction and the alignment of peaks. In our opinion, for the set of chromatographic fingerprints being studied, the improvement of the signal-to-noise ratio mostly required the elimination of the heteroscedastic component of signal noise. Its influence was diminished using a simple logarithmic transformation (log10) that was proposed by [8].

Furthermore, the preprocessing of chromatographic fingerprints consisted of the removal of irregular baseline components using the penalized asymmetric least squares approach. Although input parameters can be tuned for each chromatographic fingerprint individually in PALS, satisfactory results were obtained for the following settings of input parameters: $d = 2$, $p = 10^{-4}$ and $\lambda = 10^4$.

Once the signal-to-noise ratio of the chromatographic fingerprints was improved, they were aligned using the correlation optimized warping method. Prior to the alignment of signals, a target signal was selected as described in reference [28]. The selected target signal was considered to be the most representative with respect to all of the chromatograms. It was found to be the chromatogram that represented a group of samples in which TBT was not detected. On the other hand, no characteristic peak from the TBT component was present in the region that is relevant for the TBT identification along elution time axis in this target signal. In the course of the alignment procedure, improved alignment could be observed for the peaks that were located outside the elution time region where analyte of interest was present. It was impossible to conclude beforehand that such a target candidate would complicate the alignment task. To study the possible enhancement of the alignment performance, another target signal was selected using the same principle, but this time from the group of samples that contained TBT.

In general, two different target signals and various settings of input parameters (section length and the slack parameter) were tested to examine the effect of alignment. From all of the settings of input parameters, results for two extreme pairs of settings are discussed in detail — a short section (with 20 sampling points) and a longer section (with 28 sampling points), as well as a small slack value (2) and larger value (4). Bearing in mind the total number of sampling points, considered lengths of sections result in warping of ca. 102 and 143 sections, respectively.

Two histograms that depict the distribution of the correlation coefficients, which were calculated between all of the chromatographic fingerprints and a particular target signal that was selected either from a whole set of signals or from the TBT group only, are presented in Fig. 2. It becomes apparent, that in general, chromatographic signals are very diverse and their similarity with respect to any target signal is relatively small. The histogram that is presented in Fig. 2a indicates that more than 450 chromatographic fingerprints have a correlation coefficient above 0.4 with respect to the target signal. When a target chromatogram is selected from a group of chromatograms that confirms the presence of TBT (see Fig. 2b), a substantially smaller number of chromatograms have their initial correlation with the target chromatogram above a value of 0.4.

To compare the results of the alignment with respect to the selected input parameters, the differences between the initial and the final correlation coefficients that were calculated between a target signal and the aligned signals were monitored. The sum of all of these differences can be considered as a simple measure of the alignment performance and should be maximal. We will refer to this measure as the overall alignment gain (AG). Apparently, regardless of the settings of input parameters and target variant that are being considered, the values of the correlation coefficients increased after the alignment using COW. The best alignment results were obtained when the warped sections contained twenty sampling points and the slack parameter was set to four. In this case, the overall score of the alignment gain was equal to 114.5 after the alignment.

More detailed information regarding the input parameters and corresponding alignment improvement is provided in Table 1.

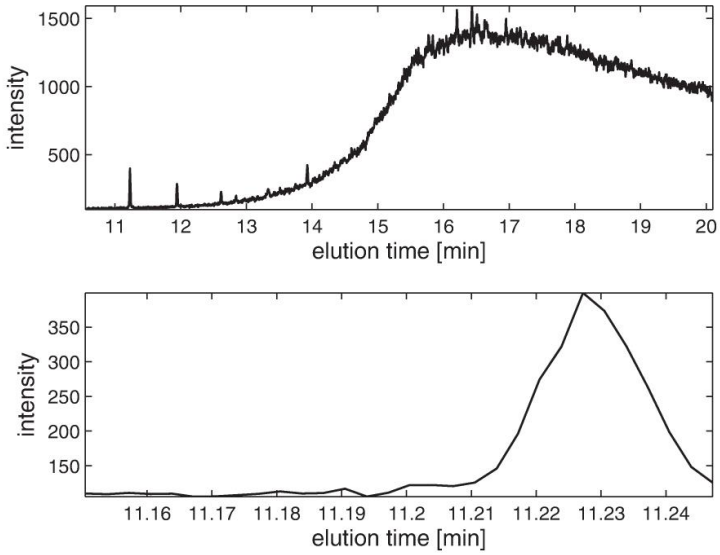


Fig. 1. Example chromatogram of a water sample containing tributyltin with the enlarged region containing peak of analyte.

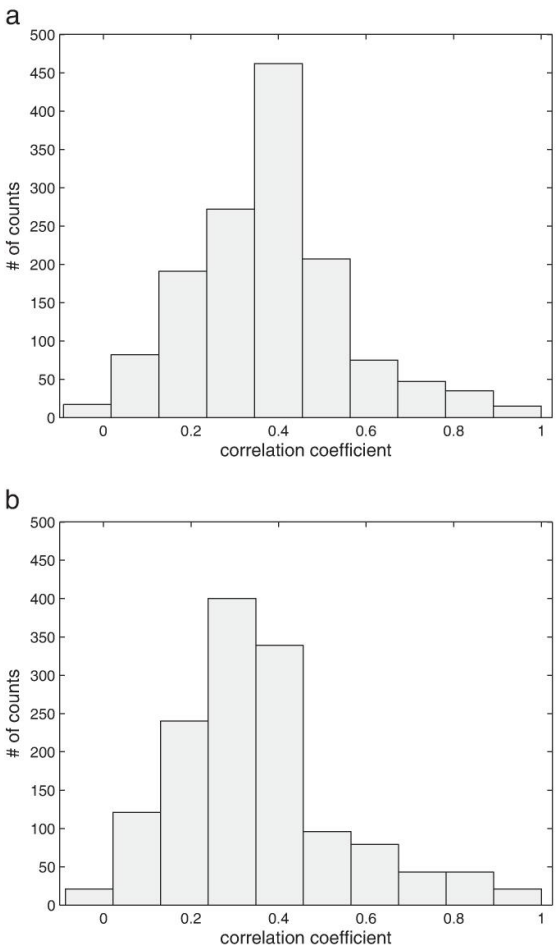


Fig. 2. Histograms of the correlation coefficients that were calculated between the chromatographic fingerprints and a target signal selected from a group of samples: a) without TBT and b) with TBT.

3.2. Studying the effect of preprocessing upon model's performance

In the next step, several discriminant models, which were based on the PLS-DA approach, were constructed in order to study the effects of signal preprocessing upon a model's performance as well as the possibility of the inferring presence of TBT directly from the chromatographic fingerprints in detail.

Table 1
Results of the alignment of signals using the COW method for a given section length (N) and slack value (t), which is expressed as the overall gain (AG).

Target	No.	N	t	$\Sigma\Delta(+)$	$\Sigma\Delta(-)$	AG
Selected from the TBT free group	1	28	2	+62.5717	−2.8840	+59.6877
	2	28	4	+78.2420	−3.0301	+75.2119
	3	20	2	+73.0242	−2.5966	+70.4276
	4	20	4	+107.5005	−4.5093	+102.9912
Selected from the TBT group	1	28	2	+80.0706	−2.5456	+77.5250
	2	28	4	+103.6455	−3.5887	+100.0568
	3	20	2	+89.8385	−2.9274	+86.9111
	4	20	4	+119.2912	−4.7873	+114.5039

To unify the further presentation of the results that were obtained from different models, all of the models were constructed using the same scheme for data containing two groups of samples — with TBT (157 samples) and without TBT (1246 samples). Namely, they were built in the course of the bootstrap procedure, which was repeated 500 times (without replacement) in order to (i) permit the construction of a balanced model and training sets, (ii) provide a good model validation framework and (iii) obtain information about the uncertainty level for the figures of merit that were being monitored. For each bootstrap sample, a model set was constructed by randomly drawing 79 samples from each group to form the model set and 78 samples from the remaining samples to form the test set. The model and test sets were exclusive.

For a given bootstrap sample, ten PLS-DA models were constructed with f = 1, 2,..., 10 PLS factors. Discriminant performance of the PLS-DA models is visualized by plotting the mean values of the correct discrimination rates, which are obtained independently for the model and test sets in the course of the bootstrap procedure, in the function of the number of PLS factors. For each mean value of the correct discrimination rate, its corresponding standard deviation, which was estimated from 500 bootstrap values, is indicated as a vertical error bar.

At this point, we tested whether the identification of samples that contain TBT can be supported using long chromatographic fingerprints (elution time from 10.55 min up to 20.10 min) would be preferred. In general, we expect that a good discriminant method should be able to pick up any relevant chromatographic features (elution time regions) during the construction of a model. Moreover, the presence or absence of TBT in environmental samples can potentially be correlated with the presence or absence of other chemical substance(s) and thus may potentially reinforce discrimination. Bearing in mind that chromatographic conditions are virtually the same for each sample, a specific pattern or profile can also be helpful in improving a discriminant rule and increasing the flexibility of a model with respect to new sources of variation.

Consecutive sub-plots of Fig. 3 illustrate the performance of the PLS-DA models constructed for the raw chromatographic fingerprints and the chromatographic fingerprints preprocessed that were differently (after baseline elimination, square root transformation, log10 transformation and alignment or after using a combination of selected preprocessing techniques).

A careful analysis of the PLS-DA results indicated a few interesting observations. For the raw chromatographic fingerprints, the relatively poor performance of the PLS-DA model improved when a larger number of latent factors were used for its construction. Unfortunately, simple models, which include a small number of latent factors (e.g., three), offer an error rate of predictions below 0.7. On the other hand, a relatively large complexity of the model can be explained as an attempt to compensate for the error structure and vague or imprecise information being present in the data that is related to the TBT content. Taking into account the presence of the substantial heteroscedastic noise component in chromatographic fingerprints, the effect of two different signal transformations was tested on the PLS-DA performance — the square root and the log10 transformation. In general, they led to a slight improvement of the model's predictions and unified the error levels over the model's complexity domain. The best improvement in the predictions, which was observed for log10 transformed chromatographic fingerprints, exceeded a 0.7 level of a correct classification rate for a two-factor PLS-DA model. Removing the baseline component from the chromatographic fingerprints improved the performance of the discrimination models. The correct classification rate systematically increased along with the complexity of the model. Compared to the models that were obtained for the raw chromatographic fingerprint with four component model, a nearly 0.75 of the correct recognition of the test set samples was achieved. The largest improvement in prediction performance was observed when the chromatographic fingerprints were baseline corrected and the effect of the heteroscedastic noise was suppressed using either the square root or the log10 transform. After

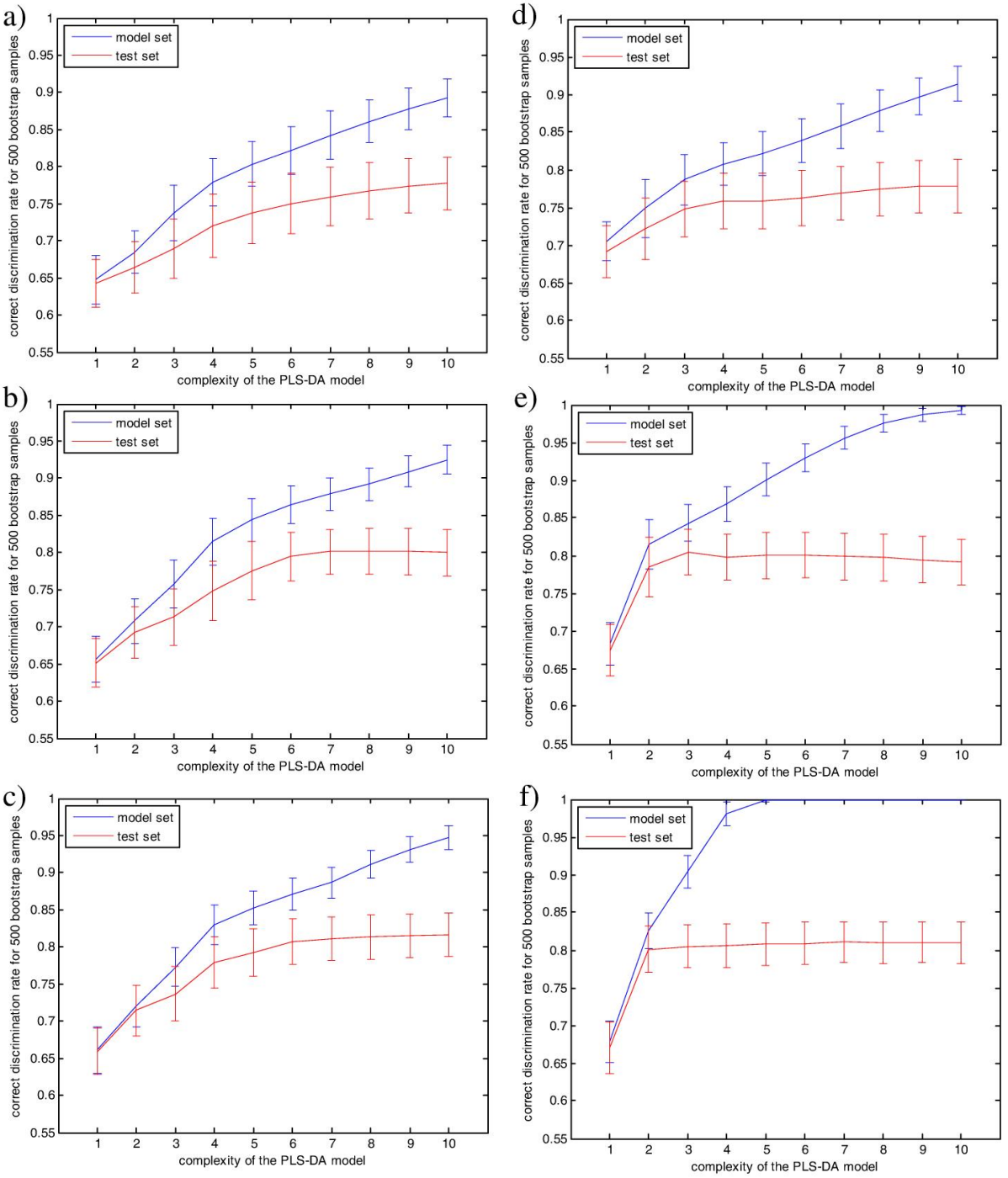


Fig. 3. Performance of the PLS-DA models constructed for a) row chromatographic fingerprints and those that were differently preprocessed, b) square root transformation, c) log10 transformation, d) baseline corrected, e) baseline corrected and square root transformed and f) baseline corrected and log10 transformed.

applying the former transformation, the three-component PLS-DA model had a 0.8 correct classification rate for the test set samples, whereas its sensitivity and specificity were equal to 0.820 and 0.702, respectively.

3.3. Modeling based on relevant chromatographic regions

In the next step, a variable selection approach, which was based on the selectivity ratio filter, was implemented within PLS-DA in the course of the bootstrap procedure. This was done in order to identify the relevant chromatographic regions along the elution time axis. The mean values of the selectivity ratios for each variable were obtained as described in [29] and [30] and are presented in Fig. 4a. In general, variables with selectivity ratios above one can be considered as relevant for the problem that is being studied. In Fig. 4a the two sharp peaks that are located between 11 and 12 min indicate the groups of the relevant variables. In fact, within this retention time window, a chromatographic peak of TBT analyte was present.

When the PLS-DA model was built for a reduced number of variables that were located in two narrow elution time windows that contained ten sampling points each with centers in the maxima of those two peaks with a three-component model, a nearly 0.82 correct classification rate was achieved (see Fig. 4b). Without any loss of model performance, similar discrimination results were obtained, but for a very limited number of variables (only twenty variables).

The percentages of incorrectly identified water samples based on the discriminant models are indicated for each group of samples separately in Fig. 4c. It is interesting that similar error rates approaching ca. 0.19 were obtained for both groups of samples and that their uncertainty levels were comparable. Compared to the PLS-DA model that was built for complete chromatographic signals, the model built for a limited number of variables had similar values of sensitivity and specificity, which were equal to 0.821 and 0.726, respectively.

In order to construct additional models, input parameters for LK-SVM, KNN (the number of neighbors, $k = 9$), LR (the α parameter has been cross validated) and PC (width of window, $\sigma = 0.3$) were optimized during the cross-validation procedure. AB and RF models were constructed based on 100 decision trees that had two terminal nodes.

The results that were obtained from eight different models for raw chromatographic fingerprints favored the use of a relatively simple linear PLS-DA model. Apparently, the construction of a regularized model (LR), ensemble models (RF and AB), maximizing the group margins (LK-SVM) or a non-linear boundary (KNN and PC) did not substantially improve the figures of merit that were obtained (the correct classification rate, sensitivity and specificity) for the test set samples. For all of the models, the level of uncertainty that was associated with the estimates of their sensitivity and specificity was relatively small (below 0.066).

In the next step, a PLS-DA model was constructed for the preprocessed chromatographic fingerprints (baseline corrected, aligned and log10 transformed). A target signal was chosen from group of samples that contained TBT for fingerprint alignment. The input parameters for COW were selected to guarantee to the greatest degree of alignment improvement (see Table 1). After the alignment, the chromatographic fingerprints were transformed using the log10 transformation. The correct discrimination rate approached 0.80 with the sensitivity and specificity equal to 0.800 ± 0.048 and 0.790 ± 0.054, respectively. Virtually

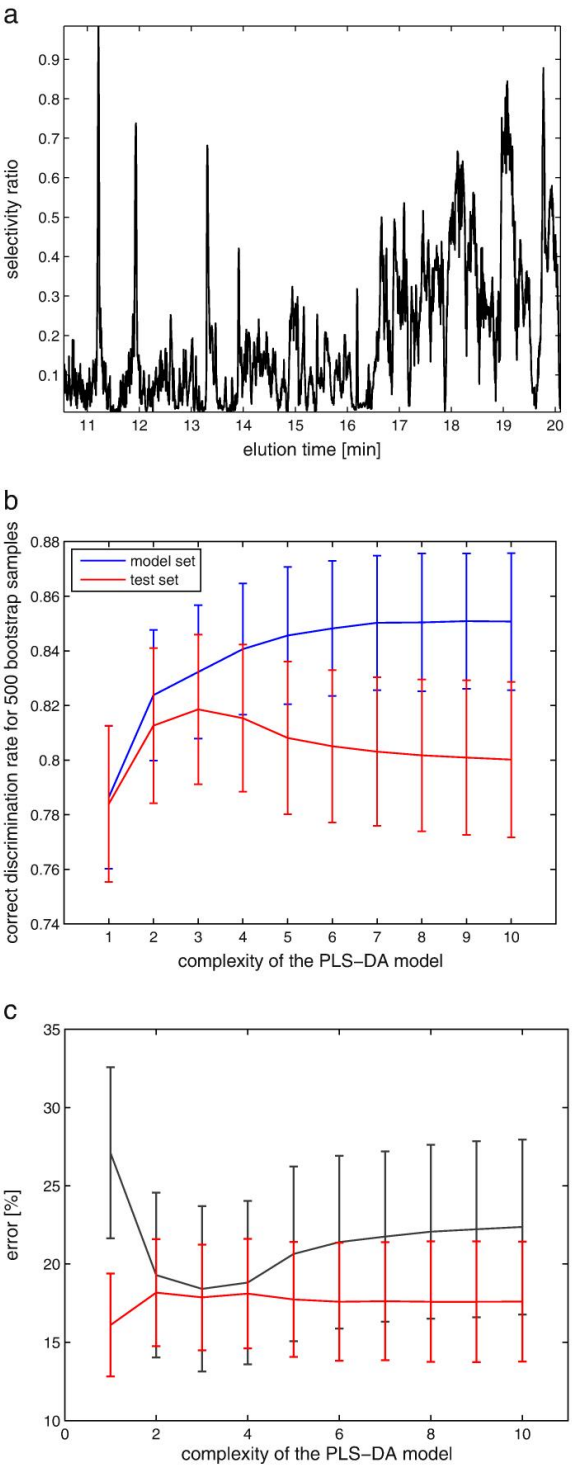


Fig. 4. a) Values of the selectivity ratio for different variables, b) performance of the PLS-DA model that was constructed for a subset of variables (30 variables with selectivity ratio values above one) expressed as the correct classification rate for the training and test set samples together in a function of the number of latent PLS factors (with uncertainty estimated during the bootstrap method) and c) percentages of the error rates in the function of model complexity that was obtained for each group of water samples (with uncertainty estimated during the bootstrap method).

the same predictions in terms of the correct classification rate were obtained from the PLS-DA model that was built for a reduced number of chromatographic features selected with average selectivity ratio above 0.67 using the SR filter (in total 67 variables). It is apparent from the data in Table 2 that most of models helped to achieve correct classification rates that were very close to 0.80. However, a balanced model performance with respect to similar sensitivity and selectivity is offered by the AB model, which can be considered to be the best one. Virtually the same discriminant performance can be obtained from the RF model. However, the differences between classic PLS-DA and LR, LK-SVM are relatively small and the results are not considerably worse than the best model.

From the results that were obtained, one can conclude that nonparametric and local models such as KNN and PC, which offer a nonlinear advantage, do not outperform simple linear models. Moreover, the PC model has a considerably low performance for all of the chromatographic fingerprints (raw and preprocessed) in terms of figures of merit that were considered. This finding is not surprising. By definition, the performance of PC depends on the selection of the relevant variables (region(s) along the elution time axis) either by using expert knowledge or a variable selection technique.

3.4. Models built for imbalanced training sets

Thus far, we have discussed the results that were obtained from different models that were built and validated using the balanced training and testing sets of samples, i.e. including the same number of positive and negative samples. Such a philosophy has been driven by the following two facts: (i) the performance of some discriminant models, including PLS-DA, is affected when the training set is unbalanced and (ii) when a test set is balanced, and in the same manner the figures of merit that describe the performance of a model for both groups of samples can be verified.

However, this is not the case in many real-life situations when, by definition, the proportion between the number of positive and negative samples is very small. For instance, in our study, the data set that was collected contained 157 positive and 1246 negative water samples in which the presence or absence of TBT was confirmed in accordance with the PN-EN ISO 17353:2006 procedure.

Analysis of the chromatograms that described the water samples with a confirmed presence of TBT indicated that the elution time window in which the analyte that was of interest should elute was relatively narrow (11.15–11.25 min). Therefore, the maximal effective time for chromatographic analysis is less than 12 min. Afterwards, for the next 8 min, the chromatograms were developed in order to remove the residual substances that were retained by a chromatographic column.

If we take into account the effective range of the elution time, the large proportion of negative samples and the simplicity of data treatment (no extensive preprocessing), the results that were obtained from the six models are very encouraging. During model construction, the bootstrap approach was used. Training sets were randomly drawn and they contained 78 positive and 623 negative samples, whereas test sets contained 79 positive and 623 negative samples. From the theoretical point of view, except for the PLS-DA approach, the remaining models were not affected by the uneven proportion of samples from the two groups that were included in the training set.

Prior to the construction of the models, the input data (chromatographic signals within the range 11.15–11.25 min) were preprocessed using the $\log10(x_{n+1}/x_n)$ transformation in order to remove any possible closure effect. The following settings were found to be optimal for the different models: LR ($\lambda = 1$), LK-SVM ($C = 1$), RF and AB (50 trees, 4 terminal nodes), KNN ($K = 11$), PC ($\sigma = 1$). Regardless of the type of model that was being considered, they all had excellent performance in terms of their specificities — larger than 0.930 (the smallest value was observed for LR). Therefore, only a few water samples were incorrectly recognized as containing TBT by a model. It is also worth noting that the levels of uncertainty that were associated with the estimation of specificity values were very small and thus the estimates that were obtained can be considered to be highly accurate. As can be seen from Table 3, the best model is offered by the RF approach and is characterized by a specificity value equal to 0.989 ± 0.005, which means that one can claim that a water sample does not contain TBT when it is recognized as TBT free with a 0.98 probability using the model that was constructed.

4. Conclusions

The application of different classifiers to a database of chromatographic fingerprints of environmental water samples, as illustrated in this study, is in fact an attempt to build a reliable expert system using machine learning methods. It is interesting to note, that even for the raw chromatographic fingerprints that were obtained in the course of the routine monitoring of the TBT contaminant in environmental water samples, i.e. without any preprocessing, the success of its detection is quite promising. This seems to be confirmation of the hypothesis that relevant information regarding the presence of TBT is expressed by chromatographic fingerprints. For chromatographic fingerprints that were modeled, their overall quality influenced the performance of different classifiers to a lesser extent. Building an expert system based on the available collection of chromatographic fingerprints, which represents different sources of variation, seems to be a very appealing approach useful in a testing laboratory. It can be used as an additional support for decision making, and thus reducing the time that is required

Table 2
Results that were obtained from the models expressed as the correct classification rate (CCR), sensitivity and specificity of the models for the samples from independent test set. Models no. 1–8 were constructed for raw (R) chromatograms, whereas models no. 9–16 were constructed for preprocessed chromatograms (P) — after baseline correction, log10 transformation and alignment of signals. Estimates of the uncertainty that was associated with the mean values of the correct classification rate (CCR), sensitivities and specificities for a given model were obtained using the bootstrap procedure (500 bootstrap samples). Asterisk denotes results obtained from PLS-DA extended with variable selection approach (the selectivity ratio).

Data	No.	Model	CCR	Sensitivity	Specificity
R	1	PLS-DA	0.805 ± 0.028	0.819 ± 0.046	0.791 ± 0.059
	2	PLS-DA*	0.813 ± 0.028	0.818 ± 0.034	0.807 ± 0.053
	3	LR (L^2)	0.790 ± 0.032	0.798 ± 0.050	0.782 ± 0.054
	4	LK-SVM	0.795 ± 0.031	0.813 ± 0.047	0.785 ± 0.051
	5	RF	0.767 ± 0.035	0.793 ± 0.055	0.742 ± 0.063
	6	AB	0.783 ± 0.033	0.799 ± 0.055	0.766 ± 0.063
	7	KNN	0.679 ± 0.035	0.744 ± 0.066	0.614 ± 0.064
	8	PC	0.668 ± 0.030	0.416 ± 0.055	0.919 ± 0.033
P	9	PLS-DA	0.795 ± 0.030	0.800 ± 0.048	0.790 ± 0.054
	10	PLS-DA*	0.793 ± 0.029	0.841 ± 0.038	0.655 ± 0.057
	11	LR (L^2)	0.778 ± 0.035	0.781 ± 0.059	0.775 ± 0.053
	12	LK-SVM	0.781 ± 0.029	0.784 ± 0.055	0.774 ± 0.053
	13	RF	0.798 ± 0.036	0.800 ± 0.064	0.793 ± 0.067
	14	AB	0.806 ± 0.035	0.801 ± 0.055	0.813 ± 0.066
	15	KNN	0.718 ± 0.036	0.854 ± 0.049	0.581 ± 0.087
	16	PC	0.581 ± 0.031	0.289 ± 0.058	0.873 ± 0.050

Table 3
Results that were obtained from the different models using imbalanced training sets, which are expressed as the correct classification rate (CCR), sensitivity and specificity along with their levels of uncertainty, which was estimated based on the Monte Carlo approach (500 bootstrap samples).

No.	Model	CCR	Sensitivity	Specificity
1	LR (L^2)	0.904 ± 0.012	0.699 ± 0.066	0.930 ± 0.014
2	LK-SVM	0.921 ± 0.012	0.430 ± 0.072	0.983 ± 0.007
3	RF	0.927 ± 0.012	0.439 ± 0.063	0.989 ± 0.005
4	AB	0.922 ± 0.012	0.422 ± 0.067	0.985 ± 0.006
5	KNN	0.931 ± 0.011	0.512 ± 0.063	0.984 ± 0.006
6	PC	0.923 ± 0.011	0.458 ± 0.065	0.982 ± 0.006

to complete analytical workflow and thus increase the throughput of a testing laboratory. The relatively high prediction performance of the models that were constructed (with a correct classification rate approaching 0.8, and sensitivities and specificities about 0.8) indicates that fingerprints of environmental water samples with and without TBT can serve as a database of water samples. Its modeling opens a unique possibility for the construction of an expert system that is based on logic rules that can be built using certain machine learning methods. In any case, the function of the purpose should be taken into account when optimizing the analytical procedures. Based on the database collection and constructed models, if a sample of Polish inland water that is analyzed in accordance with PN-EN ISO 17353:2006 is recognized using a model as TBT, it can be claimed with at least a 0.930 ± 0.014 probability (logistic regression model) and 0.98 using the remaining models. These results provide considerable evidence that the database that is available contains relevant information and shows the relatively large diversity of Polish inland water samples. Obviously, the analytical personnel of any testing laboratory during routine/high-throughput analysis cannot neglect calibration procedures, but in fact qualitative information will be stored in the database. Therefore, they can not only use a single reference chromatogram efficiently, but the entire collection. A major advantage of such a database relies on available at a given moment of time knowledge of analytical staff inferred from chromatograms on one hand and analytical decisions on the other. It should be borne in mind that in this context 'decision making' means the result of any procedure that is carried out by an analytical chemist that reflects his/her experience gained over time, the results of analysis that are obtained, the analysis of replicated measurements results, etc.

Moreover, such databases and expert systems can also be built for the routine monitoring of other priority substances in different media, e.g., excise duty components in diesel fuel.

Conflict of interest

The authors declare that there are no conflicts of interest.

Acknowledgments

MD wishes to acknowledge the support of the National Science Centre, Poland (research grant no. 2014/13/B/ST4/05007).

BK is grateful for the financial support within the framework of the DoktoRIS program — scholarship program for innovative Silesia co-financed by the European Union under the European Social Fund.

The authors would like to thank Chief Inspectorate of Environmental Protection in Warsaw, Poland for permission to use in this study results of chromatographic analysis of water samples.

References

[1] M. Daszykowski, B. Walczak, Use and abuse of chemometrics in chromatography, TrAC Trends Anal. Chem. 25 (2006) 1081–1096, http://dx.doi.org/10.1016/j.trac.2006.09.001.
[2] J.M. Amigo, T. Skov, R. Bro, ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics, Chem. Rev. 110 (2010) 4582–4605, http://dx.doi.org/10.1021/cr900394n.
[3] A.C. Duarte, S. Capelo, Application of chemometrics in separation science, J. Liq. Chromatogr. Relat. Technol. 29 (2006) 1143–1176, http://dx.doi.org/10.1080/10826070600574929.
[4] B.J. Stojanovič, Y. Dotsikas, Chemometrics in chromatography, Chromatographia 76 (2013) 207–209, http://dx.doi.org/10.1007/s10337-013-2401-2.
[5] R. de Carvalho Oliveira, R.E. Santelli, Occurrence and chemical speciation analysis of organotin compounds in the environment: a review, Talanta 82 (2010) 9–24, http://dx.doi.org/10.1016/j.talanta.2010.04.046.
[6] M. Bravo, G. Lespes, I.D. Gregori, H. Pinochet, M.P. Gautier, Determination of organotin compounds by headspace solid-phase microextraction–gas chromatography–pulsed flame-photometric detection (HS-SPME–GC–PFPD), Anal. Bioanal. Chem. 383 (2005) 1082–1089, http://dx.doi.org/10.1007/s00216-005-0131-5.
[7] M.M. Bravo, L.F. Aguilar, W.V. Quiroz, A.C. Olivieri, G.M. Escandar, Determination of tributyltin at parts-per-trillion levels in natural waters by second-order multivariate calibration and fluorescence spectroscopy, Microchem. J. 106 (2013) 95–101, http://dx.doi.org/10.1016/j.microc.2012.05.013.
[8] O.M. Kvalheim, F. Brakstad, Y. Liang, Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise, Anal. Chem. 66 (1994) 43–51, http://dx.doi.org/10.1021/ac00073a010.
[9] P.H.C. Eilers, A perfect smoother, Anal. Chem. 75 (2003) 3631–3636, http://dx.doi.org/10.1021/ac034173t.
[10] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, J. Chromatogr. A 805 (1998) 17–35.
[11] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemometrics 17 (2003) 166–173, http://dx.doi.org/10.1002/cem.785.
[12] E.K. Kemsley, Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods, Chemom. Intell. Lab. Syst. 33 (1996) 47–61, http://dx.doi.org/10.1016/0169-7439(95)00090-9.
[13] P.M. Williams, Bayesian regularisation and pruning using a Laplace prior, Neural Comput. 7 (1994) 117–143.
[14] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, J. Royal Stat. Soc. Ser. B 67 (2005) 91–108, http://dx.doi.org/10.1111/j.1467-9868.2005.00490.x (Statistical Methodology).
[15] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Royal Stat. Soc. Ser. B 67 (2005) 301–320.
[16] V.N. Vapnik, Statistical Learning Theory, 1 edition Wiley, New York, 1998.
[17] T. Czekaj, W. Wu, B. Walczak, About kernel latent variable approaches and SVM, J. Chemom. 19 (2005) 341–354.
[18] B. Walczak, D.L. Massart, The radial basis functions — partial least squares approach as a flexible non-linear regression technique, Anal. Chim. Acta 331 (1996) 177–185, http://dx.doi.org/10.16/0003-2670(96)00202-4.
[19] Y. Freund, R.E. Schapire, A Short Introduction to Boosting, 1999.
[20] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.
[21] A.Y. Ng, Feature selection, L1 vs. L2 regularization, and rotational invariance, Proceedings of the Twenty-First International Conference on Machine Learning, ACM, New York, NY, USA 2004, p. 78, http://dx.doi.org/10.1145/1015330.1015435.
[22] R. Gosselin, D. Rodrigue, C. Duchesne, A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications, Chemom. Intell. Lab. Syst. 100 (2010) 12–21, http://dx.doi.org/10.1016/j.chemolab.2009.09.005.
[23] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.-M. Myhr, O.M. Kvalheim, Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles, Anal. Chem. 81 (2009) 2581–2590, http://dx.doi.org/10.1021/ac802514y.
[24] T.N. Tran, N.L. Afanador, L.M.C. Buydens, L. Blanchet, Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC), Chemom. Intell. Lab. Syst. 138 (2014) 153–160, http://dx.doi.org/10.1016/j.chemolab.2014.08.005.
[25] R. Wehrens, H. Putter, L.M.C. Buydens, The bootstrap: a tutorial, Chemom. Intell. Lab. Syst. 54 (2000) 35–52, http://dx.doi.org/10.1016/S0169-7439(00)00102-7.
[26] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[28] M. Daszykowski, B. Walczak, Target selection for alignment of chromatographic signals obtained using monochannel detectors, J. Chromatogr. A 1176 (2007) 1–11, http://dx.doi.org/10.1016/j.chroma.2007.10.099.
[29] S. Deja, I. Porebska, A. Kowal, A. Zabek, W. Barg, K. Pawelczyk, et al., Metabolomics provide new insights on lung cancer staging and discrimination from chronic obstructive pulmonary disease, J. Pharm. Biomed. Anal. 100 (2014) 369–380, http://dx.doi.org/10.1016/j.jpba.2014.08.020.
[30] B. Krakowska, I. Stanimirova, J. Orzel, M. Daszykowski, I. Grabowski, G. Zaleszczyk, et al., Detection of discoloration in diesel fuel based on gas chromatographic fingerprints, Anal. Bioanal. Chem. 1–12 (2015), http://dx.doi.org/10.1007/s00216-014-8332-4.

dr hab. Michał Daszykowski, prof. UŚ
Instytut Chemii
Uniwersytet Śląski
ul. Szkolna 9
40-006 Katowice

Katowice 29.06.2016

Oświadczam, że w artykule pt. „Expert system for monitoring the tributyltin content in inland water samples" opublikowanym w czasopiśmie Chemometrics and Intelligent Laboratory Systems, 149 (2015) 123-131 mój udział polegał na:

- współtworzeniu hipotezy badawczej i ogólnej koncepcji badań,
- weryfikacji poprawnego wykorzystywania narzędzi chemometrycznych (PAsLS, COW, PLS-DA, SR),
- pomocy w interpretacji wyników,
- opiece i merytorycznym nadzorze procesu przygotowania manuskryptu,
- pomocy w przeprowadzeniu procedury redakcyjnej i przygotowaniu odpowiedzi na recenzje,
- dokonaniu ostatecznej korekty artykułu.

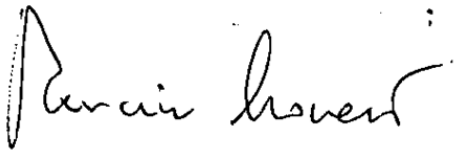
Michał Daszykowski

dr hab. inż. Marcin Korzeń
Wydział Informatyki
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie
ul. Żołnierska 49
70-210 Szczecin

Szczecin, 27.06.2016

Oświadczam, że w artykule pt. "Expert system for monitoring the tributyltin content in inland water samples" opublikowanym w czasopiśmie Chemometrics and Intelligent Laboratory Systems, 149 (2015) 123-131 mój udział polegał na:

- współpracy podczas formułowania hipotezy badawczej i ogólnej koncepcji badań,
- uzyskania wyników analizy danych chromatograficznych za pomocą wybranych metod uczenia maszynowego (LR, LKSVM, AB, RF, KNN, PC),
- pomocy w interpretacji uzyskanych wyników,
- pomocy przy tworzeniu manuskryptu i formułowaniu odpowiedzi na recenzje.



Fabiańczyk

Załącznik 4

Publikacja:	Chemometrics and identification of counterfeit medicines – a review
Autorzy:	B. Krakowska D. Custers E. Deconinck M. Daszykowski
Czasopismo:	Journal of Pharmaceutical and Biomedical Analysis
Wartość współczynnika Impact Factor	3,169

Journal of Pharmaceutical and Biomedical Analysis 127 (2016) 112–122

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com/locate/jpba

Chemometrics and the identification of counterfeit medicines—A review

B. Krakowska[a], D. Custers[b,c], E. Deconinck[b], M. Daszykowski[a,*]

[a] Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland
[b] Scientific Institute of Public Health (WIV-ISP), Operational Direction Food, Medicines and Consumer Safety, Section Medicinal Products, Rue Juliette Wytmanstraat 14, B-1050 Brussels, Belgium
[c] Research group NatuRA (Natural products and Food – Research and Analysis), Department of Pharmaceutical Sciences, University of Antwerp, Universiteitsplein 1, B-2610 Wilrijk, Belgium

ARTICLE INFO

Article history:
Received 31 December 2015
Received in revised form 31 March 2016
Accepted 14 April 2016
Available online 16 April 2016

Keywords:
Classification
Discrimination
Pattern recognition
Fingerprints
Impurity profiles
Data exploration
Data modeling

ABSTRACT

This review article provides readers with a number of actual case studies dealing with verifying the authenticity of selected medicines supported by different chemometric approaches. In particular, a general data processing workflow is discussed with the major emphasis on the most frequently selected instrumental techniques to characterize drug samples and the chemometric methods being used to explore and/or model the analytical data. However, further discussion is limited to a situation in which the collected data describes two groups of drug samples – authentic ones and counterfeits.

1. Introduction

Counterfeit medicines pose a serious threat to public health [1]. In recent years a significant increase in the number of medicine counterfeiting cases has been observed. This can be explained by easier access to modern technologies that can be used to 'copy' authentic medicines on the one hand and a lack of effective control over the medicines that are introduced into a market by different illicit vendors via internet platforms on the other hand [2]. It is impossible to obtain precise estimates of the scale of drug counterfeiting. It can roughly be expected that ca. 10% of medicines enter the worldwide market as counterfeits. On a local scale, of course, the amount of detected counterfeit medicines may differ due to certain local factors (e.g. strict and less strict legal regulations). In highly developed countries, counterfeit medicines account for ca. 1% of the total controlled market. However, more than 50% of medicines purchased over the internet through sites that disguise their physical identity are fake or poor quality drugs. As discussed in reference [3], among the most common counterfeited medicines are antimicrobials (28%), hormones (22%), antihistamines (17%), vasodilators (7%), drugs for erectile dysfunction (5%) and anticonvulsants (2%). Bearing in mind the serious consequences and dangers related to counterfeit medicines as well as the steadily growing migration of medicines all over the world, new, relatively simple and effective methods that can support the detection of counterfeit medicines as well as certain strategies are strongly desired.

Authentic medicines can usually be distinguished from counterfeits by a careful analysis of their chemical composition [1]. The most important drug ingredients are the so-called active pharmaceutical ingredients (APIs). With respect to concentration of APIs, there are four groups of medicines: (i) medicines with the correct API content and appropriate dosage, (ii) medicines with the correct API content but inappropriate dosage, (iii) medicines with an incorrect API content and (iv) medicines without any APIs (placebos).

In many cases, identifying the authenticity of a drug based on the dose and type of APIs is insufficient. The presence of impurities may alter the expected pharmacological effect of a medicine and/or increase its toxicity considerably. In general, unexpected substances are present in samples as a result of poorly controlled manufacturing conditions, the use of low-quality substrates, APIs that are produced through a different process than the certified synthesis pathway, etc. The simplest approach to testing a drug's authenticity focuses on the assessment of a specific manufacturer tags, which are deliberately introduced in order to protect

* Corresponding author.
E-mail address: michal.daszykowski@us.edu.pl (M. Daszykowski).

http://dx.doi.org/10.1016/j.jpba.2016.04.016
0731-7085/

a product. These unique chemical or visual tags can be related to type/composition of the applied packing materials, unique holograms, labels imprinted on the surface of a tablet, etc. Another group of approaches relies on the characterization of the chemical composition of drugs. These concentrate on the analysis of the API content and, if necessary, a determination of a complete chemical profile. The API content that is found using a given analytical approach is then compared with the one declared by the manufacturer. Such a comparison is carried out by testing the null hypothesis postulating that there is no significant difference between the composition of the tested sample and the declared API level(s) using the *t*-test. Accepting the null hypothesis implies that a tested sample is authentic with respect to its API level. However, in most cases, determination of an API is insufficient to confirm authenticity. An alternative approach assumes that medicine samples are described by different unique instrumental signals without the need to determine the chemical content (the so-called chemical fingerprints) and/or additional parameters. Depending on the selected method, analytical data are collected with the hope that they can support the differentiation between authentic and counterfeit medicines. By definition such data are multivariate and thus their exploration and modeling requires the use of chemometric methods to extract useful chemical information.

In this review paper, we focus our attention on a presentation of the possibilities that arise from the effective use of chemometric methods in the field of drug verification, specifically those that are applied to distinguish between authentic and counterfeit drug samples. The list of chemometric approaches discussed in the consecutive chapters of this article is not exhaustive. However, it can be considered to be a good starting point for the further exploration of the chemometric toolbox. In supplementary material we provide a list of methods acronyms.

2. Analytical techniques used to describe medicine samples

Nowadays, various analytical methods are used to characterize medicines and verify their authenticity, see, e.g. [4]. In the toolbox of analytical methods one can find relatively simple ones such as the colorimetric methods, e.g. [5], dynamic thermal analysis [6] and advanced ones such as liquid chromatography (LC) [7], high-performance liquid chromatography (HPLC) [8], gas chromatography (GC) [9], capillary electrophoresis [10], mid-infrared spectroscopy and Raman spectroscopy [11], isotope ratio mass spectrometry (IRMS) [12], NMR spectrometry [13], mass spectrometry, etc. Hyphenated techniques are also often used. In general, these analytical platforms provide a user with large amounts of analytical data. A sample is usually characterized by hundreds or even thousands of measurements. This makes data exploration, modeling and interpretation very complex. Chemometric approaches can be used to deal with an excess of explanatory variables efficiently in order to extract meaningful information from the collected data. They are designed to study the possible relations and/or existing differences between different groups of samples (e.g. authentic and counterfeit medicines).

2.1. Chromatographic-based methods

Chromatographic techniques are most frequently used to determine the chemical composition of medicines [4]. The area of their application is very wide because they have the potential to separate the different components of mixtures and to deliver qualitative and quantitative information for them.

Simple thin-layer chromatography (TLC) has been effectively introduced in many laboratories whose focus is drug verification. Its popularity stems primarily from the low costs of an analysis, limited instrumental requirements and straightforward interpretation of separation results. Because of the simplicity of the TLC approach, one can find examples illustrating the use of TLC in the context of authenticity studies in the literature [14]. For instance, the detection of counterfeit Plavix® tablets can be achieved using the TLC separation of artemisinin and its derivatives followed by detection based on a color reaction [15].

High-performance liquid chromatography (HPLC) is probably the most popular instrumental chromatographic technique for the analysis of pharmaceuticals. It is regarded as a reference method in the qualitative and quantitative analysis of many pharmaceutical substances and also serves as a reference method in the validation of a large number of analytical techniques. HPLC systems can be equipped with various types of detectors that offer interesting detection properties, for instance, mass spectrometry (MS), diode-array detector (DAD) and evaporative light scattering (ELS), which help to increase the sensitivity, accuracy and precision of a method. These features are strongly desired especially when the analysis is focused on the chemical components that are present in a sample at low concentrations (e.g. impurities).

Gas chromatography (GC) is typically used for the analysis of volatile substances that are stable at high temperatures. Like HPLC, GC is accurate and repeatable and its sensitivity is determined by the detector that is (the type of mass spectrometer or flame ionization detection). In the context of drug verification, it can be used to determine the active substances, residual solvents and/or volatile impurities e.g. [9,16]. Since only a few medicines contain volatile components, the number of GC applications is smaller compared to the number of applications of HPLC or TLC.

2.2. Spectroscopic-based methods

Spectroscopic techniques operate at different ranges (energy) of electromagnetic radiation. Among the various methods, in the field of drug verification and analysis, applications of near infrared spectroscopy (NIR) [17], mid-IR analysis, FT-IR, FTIR-ATR, Raman spectroscopy [18] and nuclear magnetic resonance (NMR) [19] seem to dominate.

NIR is a spectroscopic technique that uses the near infrared region of electromagnetic radiation between ca. 700 nm and 2500 nm. It allows for the rapid analysis of samples without (or with very little) sample preparation. Spectra of the studied medicines can be obtained through packaging materials (glass or plastic blisters). Moreover, it is both non-destructive and cost-effective.

Raman spectroscopy is regarded as a versatile technique with respect to the form of a sample. It is used to analyze solid, liquid and gaseous samples. Moreover, it has the potential to perform measurements through coatings and packaging materials. Raman spectroscopy is highly selective, which allows molecules and chemical species that are structurally very similar to be identified and differentiated. Combined with chemometric methods, it has been applied in a wide range of studies focused on the identification of authentic drugs [20].

NMR spectroscopy is a frequently selected method for API verification [13], the identification of drug composition (including an analysis of impurities) [21] and for monitoring the production of medicines [22]. Moreover, the NMR spectra contain structural information that describes the sample components, which helps to determine their chemical structures. The sensitivity of this technique is insufficient for screening constituents at low concentrations as compared to the assessment of APIs (when peaks of active substances are well separated). The method is also applicable for the analysis of mixtures [19].

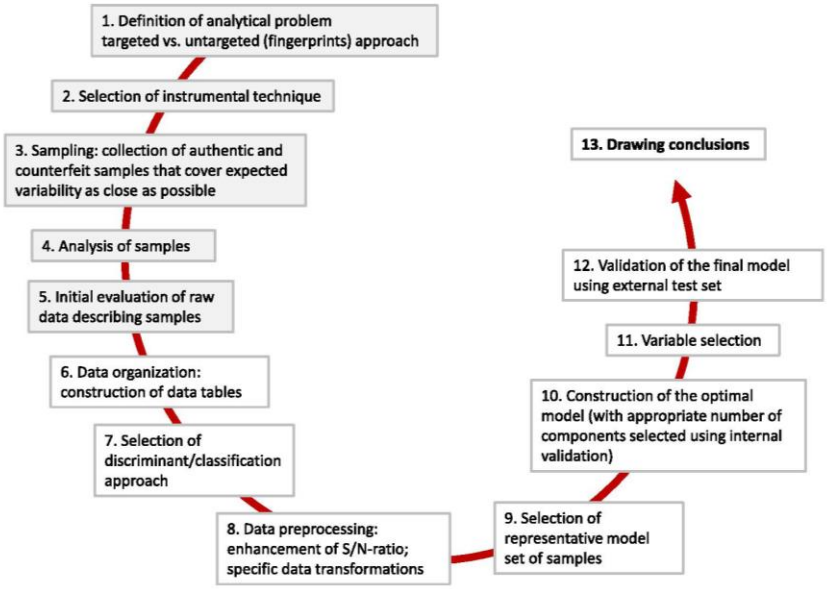


Fig. 1. The general analytical workflow that is applied in studying the authenticity of drug samples.

3. Chemometric approaches used to study the authenticity of drugs

The toolbox of available chemometric approaches is quite versatile and thus the area of their applications is large. Many methods can be directly applied to support drug analysis, including studies of the differences between authentic and counterfeit medicines.

Chemometric methods are developed with the goal of facilitating the exploration and modeling of multivariate data. Verifying the authenticity of drugs often requires a multivariate approach. This means that the differences between authentic and counterfeit samples can be observed when a sample is characterized by a comprehensive set of physico-chemical parameters. Depending on the selected analytical technique and the defined potential targets that can provide information about the authenticity of a drug, analytical data may include information about the concentrations of (i) API(s), (ii) the presence of different drug excipients, (iii) certain impurities, (iv) a collection of analytical signals (fingerprints) and/or (v) fingerprints that characterize impurities.

The general analytical workflow that is used when the authenticity of drugs is studied is summarized in Fig. 1. Steps 1–5 include experimental planning, sampling, laboratory processing and the analysis of samples, whereas steps 6–12 consists of the different stages of chemometric data treatment. The types of data and their form of representation will influence further chemometric data processing. In particular, a collection of analytical signals will require different preprocessing approaches than a set of variables that describe the concentrations of selected sample components. Once samples are analyzed, the general chemometric data processing workflow consists of several steps of equal importance. It is important to stress that these subsequent stages affect the further outcome of the chemometric analysis and thus influence the final conclusions. Firstly, data usually require some kind of preprocessing to decrease any undesired variability. Then, the preprocessed data are explored to reveal their structures. If necessary, multivariate data are further modeled using supervised methods. To gain a better understanding of the modeled phenomena and to reduce the risk of model overfitting, the relevant variables can be selected. Last, but not least, the final multivariate models must be carefully validated in order to verify their predictive power.

In general, chemometric methods can be divided into two major groups—unsupervised (the so-called exploratory approaches) and supervised techniques. Exploratory approaches help in the study of the structure of data in terms of similarities among the samples and among explanatory variables. Supervised techniques are used to construct diagnostic models that support the verification of the authenticity of a drug. The major differences between supervised and unsupervised methods are illustrated in Fig. 2.

3.1. Preprocessing of multivariate data

The aim of data preprocessing there is to improve the quality of multivariate by correcting and/or suppressing certain undesired effect(s), see reference [23] and the following chapters. These goals are achieved using mathematical transformation(s). Undesired sources of data variability 'blur' relevant chemical information and make its extraction more complex and sometimes even impossible. In general, appropriately applied data preprocessing will help

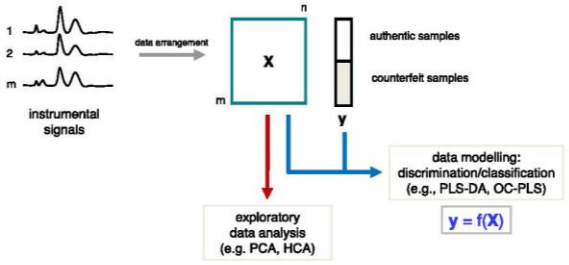


Fig. 2. Illustration of the differences between the unsupervised and supervised techniques that are applied to the collected analytical data.

to remove a large part of the data variance that is unrelated to the expected effect(s). Therefore, further data exploration and modeling will be more efficient and will lead to an easier interpretation. Certain preprocessing techniques are typically used for certain types of signals, for instance, the preprocessing of NIR spectra or the alignment of chromatographic signals.

Depending on the problem at hand, data preprocessing techniques are applied to: (i) individual explanatory variables, (ii) individual samples or (iii) to samples and variables simultaneously. Examples of the preprocessing methods that belong to these three categories are briefly discussed in the following sections of Section 3.1.

3.1.1. Preprocessing of individual explanatory variables

Transformations such as variable centering and scaling are probably the best known. They are applied independently to each explanatory variable and aim to compensate for the effect of the mean value of a variable and differences in the scales of the variables.

Centering removes a systematic difference by subtracting the mean value of a variable from each element while it preserves similarities among the data objects. A cloud of data points, which is described by the values of the measured variables (coordinates), is translated towards the origin of the coordinate system in such a way that after centering, the mean value of each variable equals zero.

Scaling transformations modify the variance of variables. In particular, scaling approaches such as standardization and z-transformation (or autoscaling) make the variance of each variable equal to one (equivalent in importance in the analysis). In addition, after the autoscaling transformation, the elements of a new variable are centered around zero.

3.1.2. Preprocessing of individual samples

This group of preprocessing methods includes (i) methods that aim to normalize samples, (ii) methods that improve the signal-to-noise ratio (denoising and background elimination), (iii) methods used to compensate for peak shifts in the analytical signals (signal alignment) and (iv) preprocessing methods that use additional information in order to remove irrelevant variations (the so-called model-based preprocessing).

In the context of chemometric data analysis, the quality of analytical signals is an important issue. Measurement errors and uncontrolled experimental or instrumental effects induce additional variability as they are the source of two additional components that decrease the quality of any analytical signal – background and noise. Depending on the selected analytical approach and instrumental method, the presence of these components requires signals preprocessing in order to draw reliable conclusions and obtain models that are characterized by good predictive abilities.

The aim of the normalization of instrumental signals is to enable a comparison of samples by the elimination of systematic bias. Essentially, normalization scales the signal elements by a constant factor. Examples of such transformations include, among others, normalization to the total area, the total sum, scaling using the Euclidean norm and the standard normal variate (SNV) [23]. It is frequently used to account for variability in the sample concentration, differences in the efficiency and sensitivity of the applied methods, variation that is observed from batch to batch and preserving biological or chemical assumptions for the system being studied. Data normalization should be applied with great caution because it results in the modification of the data correlation structure and induces spurious correlations among variables – the so-called closure effect.

Suppression of signal noise is usually achieved by signal smoothing using different filters, e.g. mean or median filters, the Whittaker smoother, smoothing with splines or Savitzky-Golay smoothing (see reference [23] and the following chapters). It is also possible to reduce signal noise by filtering out any high frequency components of a signal using the Fourier transformation or wavelets (depending on type of analytical signal).

The influence of the baseline component can be eliminated using derivatives combined with smoothing. The first derivative compensates for the effect caused by an additive type of baseline whereas the second derivative also removes the effect of a linear baseline [24]. In addition to derivatives, a baseline can be modeled using, for instance, the penalized asymmetric least squares method (PAsLS) [25].

The alignment of instrumental signals compensates for shifts in the corresponding peaks and is particularly useful when using chromatographic signals. The presence of peak shifts is usually associated with unstable experimental and/or instrumental conditions. The major objective of alignment methods is to find the optimal transformation of a signal axis (e.g. the elution time axis) with respect to a target signal. In the course of the alignment, a certain similarity measure between these two signals is maximized. Although the similarity is usually scored using the Euclidean distance, it is now more frequently scored by the correlation coefficient. Many alignment methods have been proposed in the literature, see e.g. [26]. Most of them offer a relatively large alignment flexibility and allow for the compensation of linear and non-linear peak shifts. On the other hand, large alignment flexibility due to selection of input parameters that are far from the optimal once may result in the misalignment of peaks. Dynamic time warping (DTW) was one of the early proposed alignment methods whose aim was to match the frequency spectra of words being pronounced by different speakers. Although the alignment methods are sometimes developed bearing in mind a particular type of signals, they can often be used in a wider context. The correlation optimized warping approach (COW) is a standard technique that is used to align the peaks in chromatographic signals. It can be considered to be an extension of the DTW method. Peak shifts are corrected by stretching and compressing the corresponding sections (the so-called warping) in the target signal and the signal that is being aligned. There are also methods that focus specifically on modeling the warping function. For instance, parametric time warping (PTW), semi-parametric time warping (STW) [26] and automatic alignment (AA) [27] approximate the warping function by assuming its form or spline functions. There are more examples of alignment methods in the literature, for instance, fuzzy warping, piecewise alignment, partial linear fit and genetic algorithms, which are used for peak alignment, see reference [23] and the following chapters.

Model-based preprocessing includes a family of methods that preprocess a set of explanatory variables using additional information, for instance, a response variable. Orthogonal signal correction (OSC), orthogonal projections to latent structures (OPLS) and multiplicative scatter correction (MSC) are typical examples of model-based preprocessing techniques. The aim OSC is to remove the part of data variance that is orthogonal to the modeled response. Therefore, a preprocessed set of explanatory variables contains the relevant variance for modeling. OPLS has been proposed to overcome the limitations of OSC. OPLS incorporates the OSC filter within a PLS model [28]. MSC is applied to the near-infrared spectra collected for solid samples in order to separate chemical light absorption from the physical light scattering [29].

3.1.3. Preprocessing applied to samples and variables simultaneously

In order to diminish the effect of heteroscedastic noise (the level of noise is proportional to the signal's intensity), the data element log or power transformations are used [30].

Double centering consists of subtracting the mean values from each data element row and column. The aim of this operation is to remove the closure effect. In addition to double centering, log double centering helps in the analysis of V-shaped data [31].

When samples are described by the blocks of different variables obtained from complementary instrumental techniques, scaling the data blocks to the unit variance in each block may be required prior to the analysis, see for instance reference [32].

3.2. Unsupervised methods and data exploration

There are two main groups of exploratory techniques, namely projection methods [33] and clustering methods [34]. These help to reveal any hidden structures that are present in multidimensional data. In particular, exploratory methods can provide information about any similarities that are observed among the studied samples or among the explanatory variables. The data clustering tendency can be evaluated, information about the correlation structures among variables can be obtained, variables that do not carry additional information can be identified and unique samples can be inspected based on these similarities. Exploration of multivariate data is an important step in the overall data processing workflow. Information about the existing groups of samples and/or unique objects guide the further selection of data preprocessing and data modeling methods.

3.2.1. Projection methods

Projection methods project samples that are described by a relatively large number of physico-chemical parameters into a new low-dimensional coordinate system defined by a few new variables [33]. These new variables are used to construct low-dimensional projections of the data and to visualize any hidden data structures such as groups of samples, local fluctuations in data densities and unique samples. The construction of new variables is usually realized by optimizing a certain projection index, which is defined by the goal of the chosen method. Depending on the selected projection index, new variables can enhance the modeling of the main sources of data variability and can reveal the presence of groups of samples or outlying samples. This concept was proposed by Friedman and Tukey. Interesting low-dimensional data projections are constructed in the course of the projection pursuit method (PP). The term 'interesting' usually means that the distribution of a projection is far from the normal distribution. It is scored by a projection index that is sensitive to particular data distributions, for instance, variance and entropy [33].

PP can be considered to be a general variant of a linear projection method, It can provide solutions obtained from principal components analysis (PCA), independent component analysis (ICA) and robust principal component analysis (RPCA). In certain situations, the PP method may help in the construction of more informative projections compared to the ones obtained from PCA, which is considered to be a standard approach to compress and visualize multivariate data. Low-dimensional projections of the selected principal components are constructed to maximize the variance of the projections and they tend to display interesting patterns of samples. In many situations, the expected differences among groups of samples are indeed associated with the main source(s) of data variability. In this case, score plots will indicate the groups of samples in data. Otherwise, projection indices such as entropy or kurtosis will be more efficient in capturing the data clustering tendency than the variance.

Linear projection methods seem to dominate in the majority of applications. When the compression of multivariate data with a linear projection method is ineffective, non-linear projection methods such as, e.g. Sammon's mapping, self-organizing Kohonen's maps (SOM), bottle-neck neural networks (BNN) or non-linear (kernel-based) variants of PCA are alternatives that are often considered [33].

3.2.2. Clustering methods

The objective of clustering methods is to construct groups of similar samples and/or explanatory variables [34]. Regardless of the clustering technique applied, grouping of multivariate data requires the selection of a similarity criterion. Many concepts and definitions of how to score the chemical similarity between two samples or variables have been proposed in the literature. A comprehensive overview of the available similarity measures and their properties can be found in reference [35]. The Euclidean distance, Mahalanobis distance and the correlation coefficient are among the most popular methods. In addition to the available similarity measures, the construction of groups and their linkage can be realized in a number of different ways. Bearing these options in mind, three groups of clustering methods are usually distinguished – hierarchical, non-hierarchical and density-based.

Hierarchical clustering techniques (HCA) establish a data hierarchy by the sequential grouping of the most similar samples or variables. Results of the clustering procedure are visualized by a so-called dendrogram or tree. Its bottom branches (low levels) contain the most similar samples/variables. Longer branches indicate an increase in data dissimilarity and incorporate more compact groups. Hierarchical clustering methods are primarily valued for the possibility to visualize a data structure without any assumptions about the number of clusters. The interpretation of dendrograms is very intuitive. Moreover, by first applying hierarchical clustering to samples and then to variables, one can interpret clusters of similar objects in terms of explanatory variables (the so-called two-way hierarchical clustering). Both dendrograms are extended by a color map with pixels representing particular elements of the data. The color intensity of a pixel is proportional to the corresponding measurement value.

Non-hierarchical clustering methods, also known as partitioning methods, divide data into a number of user-defined groups. In the course of the iterative procedure, they minimize a given similarity criterion. Assuming that samples from a cluster are more similar to each other than to any representative from another cluster, data partitioning is usually realized by minimizing the overall sum of the distances calculated between the center of a given group and its members. The k-means method is probably one of the most popular non-hierarchical techniques. Because they optimize such an objective function, partitioning methods perform well for clusters with spherical or elongated shapes (when Mahalanobis distance is used as a similarity measure). Non-hierarchical methods will always partition data sets into a specified number of groups. However, the constructed clusters may not correspond to the natural data clustering tendency. In addition to the k-means technique, the same objective function is used, for instance, in self-organizing Kohonen's maps, neural gas (NG) and growing neural gas (GNG) [36].

Usually, clustering results take the form of an assignment list that specifies to which cluster each object belongs (e.g. k-means, SOM, NG, GNG). In addition, it is also possible to cluster the centers of obtained groups using the HCA methods in the one- or two-way clustering mode. A color map representing the average values of the parameters for each cluster center can help in studying the major differences among groups of samples.

The clustering methods discussed are primarily used to study similarities between samples. In an exploratory analysis of drug

substances, the major focus is put on revealing differences between authentic and counterfeit medicines.

3.3. Supervised techniques—discrimination and classification

Discrimination, classification and regression methods belong to the so-called supervised data modeling techniques. Construction of such models is guided by a dependent variable(s) that specifies either to which group a sample belongs or provides information about a certain property, e.g. the concentration of API(s).

Discrimination and classification models are used to construct logic rules that help to distinguish between authentic and counterfeit medicines based on any underlying differences found at the level of their chemical composition. These techniques are also called pattern recognition methods [37]. Typical examples of discriminant methods include linear discriminant analysis (LDA), partial least squares-discriminant analysis (PLS-DA), classification and regression trees (CART) and support vector machines (SVM). Other classification models can be constructed, for instance, using the UNEQ approach, soft independent modeling of class analogies (SIMCA), classification and influence matrix analysis (M-CAIMAN), partial least squares-density modeling (PLS-DM) or the potential functions [38–40].

The major difference between classification and discrimination methods is the mechanism that determines how samples are assigned to existing groups. Classification methods construct the so-called soft classification rules that allow a sample to be assigned to one group, more than one or to any group. Discriminant methods built hard classification rules—a sample is always classified to only one group from the pool of groups being considered.

It is interesting to mention that most of the pattern recognition techniques, depending on their modifications, can perform both the discrimination and classification of samples. In general, class modeling techniques can easily be used in discrimination tasks, whereas a modification of a discriminant approach in order to derive soft classification rules is not always possible.

3.3.1. Discrimination: linear and non-linear methods

Discriminant methods divide the space of the explanatory variables into several mutually exclusive regions, the number of which is equal to the number of groups in the data. This implies that any sample is always located only in one region of the data space with respect to the values of the measured parameters—it belongs to one group.

With CART [41] the space of the explanatory variables is partitioned into a number of mutually exclusive groups of samples with respect to a dependent variable. This is achieved in the course of a recursive data partitioning process. At each step, the optimal explanatory variable and a threshold value are found that guarantee the largest decrease in the impurity function as a consequence of a binary split of the data (a logic rule). The results of CART are represented by a binary tree, which consists of nodes with samples that fulfill the logic rules that were constructed based on individual explanatory variables. Terminal nodes are the purest with respect to content of the samples and their membership. The number of terminal nodes defines the complexity of the tree and it is optimized to achieve a satisfactory predictive power of a classifier. An unknown sample is assigned to only one terminal node by evaluating the set logic rules of the tree that point to a given terminal node, i.e. the threshold values of the optimal parameters are compared with the corresponding values of the measured parameters for a sample. The logic rules that are constructed for individual explanatory variables define the regions in the data space that contain samples; however, this approach may be insufficient when the explanatory variables are correlated. On the other hand, the construction of a decision tree provides a set of relevant explanatory variables (variable selection) in a straightforward manner and offers a direct interpretation of the data structure.

LDA is used to construct linear discriminant functions, which define the directions in the space of the explanatory variables that separate the groups of samples best. These directions are found by maximizing the Fisher's criterion, i.e. the ratio between the group variance and within the group variance [37]. All of the model set samples are used to find the optimal position of the separation hyperplane. LDA has a number of assumptions that in fact limit its applicability to multivariate data—the number of samples must exceed the number of variables and the variables should be uncorrelated. It is possible to deal with the correlation among variables using PCA by replacing the explanatory variables with the corresponding principal components, which are, by definition, orthogonal.

PLS-DA is a linear discriminant variant of the classic partial least squares regression [42]. A categorical dependent variable, **y**, defines to which group a sample from a model set belongs. For a two-class discriminant problem, either bipolar ('−1' and '+1') or binary ('0' and '1') group labels are used. Once the PLS-DA model is built, the group label for a new sample is determined based on the predicted value of the dependent variable. For bipolar coding, a threshold value that assigns a sample to group '−1' or '+1' is set at 0, whereas for binary coding it is set at 0.5. Since the construction of logic rule(s) is carried out in the space of PLS factors, it is possible to deal with data that contains large number of correlated variables. Similar to LDA, all model set samples are considered to define the separation hyperplanes, but the proportions of the samples in each group affect the results of the discrimination [42].

In the k-nearest neighbor method (k-NN), the assignment of an unknown sample is based on the content of its neighborhood and is restricted to the k-nearest samples from the model set. Therefore, an unknown sample is classified to the group represented by the majority of model set samples. The distance between an unknown sample and the model set samples is scored using, for instance, the Euclidean or Mahalanobis distance. The final classification result depends on the size of the neighborhood considered, i.e. the selected number of k-nearest neighbors. The optimal number of neighbors is usually determined using a cross validation procedure. Due to its local performance, the k-NN method allows non-linear discriminant problems to be solved.

Multi-layer feedforward neural networks (MLFFNN) are composed of a number of computational units called neurons [43]. They are organized in layers (one input layer and at least one hidden layer and output layer) that are connected in a feedforward manner. Output from each neuron is computed using the so-called activation function. Depending on the selected form of the activation function, e.g. linear or non-linear, linear or non-linear discrimination tasks can be solved using MLFFNN. Training an MLFFNN is an iterative process, which is carried out using model set samples. It relies on a modification of neuron weights that usually follows the back-propagation learning approach to minimize error. The output values of the network obtained for the model set samples are compared with their group labels. Information about an error is passed back through the network in order to modify the weights of the neurons so that the error is reduced in the subsequent iterations. Although MLFFNN is a versatile technique, the training process requires a relatively large collection of samples, which have to be split into three subsets—a model set (used to train the network), a monitoring set (used to evaluate the learning process) and a test set (used to evaluate the predictive properties). Moreover, an optimal solution is not guaranteed.

The aim of support vector machines (SVM) [44], which were proposed by Vapnik, is to construct the best separation hyperplane by maximizing the margin between groups of samples. This machine learning method can be applied to solve linear and non-linear

Table 1
Examples of case studies that focus on the analysis of the most frequently counterfeited medicines extended with information about the techniques that were applied to describe the chemical composition of the samples and the chemometric methods that helped in determining the differences between authentic and counterfeit samples.

Lp.	Analyzed medicine(s)	Analytical approach used to describe the samples	Statistical and/or chemometric data treatment	Ref.
1	Antimalarial drugs	NIR spectroscopic fingerprints	PLS-DA Validation: independent randomly selected test set	[55]
2	Antimicrobial antispasmodic drugs	NIR spectra	PCA, SIMCA, multivariate image analysis Preprocessing: MSC	[3]
3	Aspirin, Paracetamol	Examination of authenticity using the isotope ^{13}C content measured with ^{13}C NMR spectrometry	PCA	[13]
4	Cialis	Analysis of packing materials using two-dimensional correlation spectroscopy	PCA Preprocessing: normalization, cosmic ray removal (algorithm of Cappel), baseline correction (algorithm of Lieber), smoothing Savitzky-Golay	
5	Cialis	Raman microscopy, API content	MCR-ALS, PCA Preprocessing: normalization, cosmic ray removal (algorithm of Cappel), baseline correction (algorithm of Lieber), smoothing Savitzky-Golay	[56]
6	Cialis	Impurity profiles (HPLC-PDA, HPLC–MS)	PCA, SIMCA, k-NN, PLS-DA Preprocessing: COW Validation: independent test set (Kennard & Stone algorithm)	[57]
7	Concor®5	NIR chemical imaging and single-point FT-NIR spectroscopy, NIR-CI	PCA, PLS Preprocessing: FT-NIR normalization of signals to unit length, EMSC; NIR-CI images: normalization of signals to unit length, smoothing with Savitzky-Golay	[58]
8	Gilbenclamide	NIR and excitation/emission fluorescence spectra	PCA, PLS-DA, SIMCA Preprocessing: Savitzky-Golay smoothing, MSC, first derivative Validation: independent test set	[59]
9	Heptodin	NIR chemical imaging and NIR spectroscopy	PCA, k-means Preprocessing: MSC, SNV, smoothing with Savitzky-Golay	[60]
10	Heptodin	NIR-CI (concentration of API)	Classical least squares, CLS Preprocessing: DPFT, dark point fixed transformation Comparison of the results with the reference method HPLC	[61]
11	Hyperglycemic drugs	Fingerprints and API concentration obtained from Raman spectroscopy	Local straight line screening, PCA Preprocessing: WT, orthogonal wavelet denoising	[62]
12	Lipitor	Raman and NIR spectra	PCA, PLS-DA Preprocessing: MSC, baseline correction by subtracting a polynomial fit Validation: independent test set (Kennard & Stone algorithm)	[63]
13	Paracetamol-containing drug	HPLC-UV impurity fingerprints	PCA, PP, HCA, GTM, auto-associative multivariate regression trees (AAMRT) Preprocessing: COW	[64]
14	Trimetoprim-Sulfametoxazole combination, Drovativarine, Metronidazole	NIR spectra	PCA, SIMCA Preprocessing: SNV Validation: independent test set	[65]
15	Viagra	Raman spectra	PCA, HCA, k-NN Preprocessing: baseline correction using ACD/Specmanager (version 9.13), smoothing Savitzky-Golay	[66]
16	Viagra	Raman microspectroscopy imaging (2D-fingerprints)	PCA, k-NN, SIMCA Preprocessing: normalization Validation: independent test set (Kennard & Stone algorithm)	[67]
17	Viagra and Cialis	Raman, NIR, FT-IR spectra	PCA, PLS Preprocessing: baseline correction; NIR spectra: SNV, FT-IR spectra: normalization using Perkin Elmer, Walham, MA, USA (version 5.0.1)	[2]
18	Viagra and Cialis	ATR-FTIR spectra	PCA, k-NN, CART, SIMCA Validation:independent test set	[68]
19	Viagra and Cialis	ATR-FTIR spectra	PCA, k-NN Validation: accuracy, sensitivity, specificity for independent test set	[69]
20	Viagra, Cialis	Chromatographic impurity profiles (HPLC-UV)	PCA, PLS-DA, k-NN, SIMCA Preprocessing: resampling, log transformation Validation: independent test set (Kennard & Stone)	[70]
21	Viagra, Cialis	Chromatographic fingerprints (HPLC-UV)	PCA, PP, HCA, k-NN, CART, SIMCA, SVM (radial basis function kernel and linear kernel) Preprocessing: resampling, autoscaling, log transformation Validation: independent test set (Kennard & Stone)	[71]
22	Viagra, Cialis	IR and Raman spectroscopic fingerprints	CART, PCA, k-NN Preprocessing: normalization, baseline correction using the Pearson's method. Validation: independent test set (Duplex algorithm)	[72]
23	Viagra, Cialis	RGB images of tables	image processing, statistical evaluation of samples distribution	[73]
24	Viagra, Cialis and their analogs	UPLC–MS and UPLC-DAD interpreted according to API concentration (Sildenafil, Tadalafil)	PCA, HCA, ANOVA	[74]
25	Viagra, Cialis and their analogs	EDX-RF spectra	PCA, HCA	[75]

discrimination tasks depending on the type of the kernel function that is applied. The widest margin is identified using a number of samples from the model set, which are called the support vectors. This requires that the constrained optimization problem, which minimizes the risk of misclassification, be solved. Assignment of a sample to a given group is based on the verification of its location with respect to the separation hyperplane. In contrast to the other pattern recognition methods discussed in Section 3.3.1, the logic rules are built locally using only a few samples that define the margin(s). Moreover, unlike MLFFNN, SVM provides a unique solution.

3.3.2. Classification—linear and non-linear methods

Classification methods, also called class modeling approaches, focus on the construction of group boundaries and classification rules for each group of samples independently [38]. Such a concept causes discriminant and classification approaches to be very different in how samples are assigned to existing groups. The soft classification paradigm, which is realized by class modeling approaches, is especially useful in a situation in which the model set samples do not represent all of the possible groups (multiclass problems). Moreover, class modeling approaches can only be used for a single group to verify whether a sample 'fits' to the group with respect to its physico-chemical parameters.

The soft independent modeling of class analogies method (SIMCA) is one of the most popular representatives of class modeling approaches. With SIMCA, each group of samples is described by the PCA model with the optimal number of principal components [37]. In general, the classifications rules are constructed based on the distances computed in the space of the principal components between an unknown sample and a given group and its residual distance to the model's space. Then, a sample is assigned to a certain group if the computed distances are shorter than a predefined threshold value.

UNEQ is another representative of classification methods [40]. In UNEQ, each group of samples is modeled using a single point that corresponds to the group centroid. The boundary of a group is represented by a contour described by the Mahalanobis distance. At a given distance (critical values of the Hotelling's T^2 statistics at a given confidence level), the probability is the same and the boundary takes an ellipsoidal form.

Recently, another group of classification approaches, the so-called one-class classifiers, has been proposed [45]. It consists of a family of classification methods that uses the partial least squares approach (OC-PLS) for the construction of logic rules. An unknown sample is classified based on the absolute centered residuals (ACR) of the response value from the optimal model and the score distance (SD) calculated between a sample and the samples from the model set in the space of the PLS factors. Corresponding threshold values for the ACR and SD distances are determined using the critical values of the normal and the F distribution (for assumed confidence levels), respectively. A sample is assigned to a given group when its ACR and SD distances are shorter than the corresponding threshold values. As demonstrated in reference [46], it is possible to derive a robust OC-PLS variant by replacing classic PLS with its robust counterpart—partial robust M-regression (PRM). Moreover, the radial basis functions-partial least squares approach (RBF-PLS) enables the modeling of groups locally, and therefore non-linear classification properties can be obtained.

The classification and influence matrix analysis method (CAIMAN) is based on the so-called influence matrix [47]. Diagonal elements of the influence matrix elements (leverages) describe the impact of each model set sample on the model's predictions. Relatively low leverage values indicate that a sample is located close to the data center, whereas large values indicate that a sample is close to a group boundary. The influence matrix is frequently used in regression analysis. Samples are assigned to a given group based on the output of the membership function and the assumed threshold value.

3.3.3. Selection of model set samples

Construction of an adequate calibration or classification/discrimination model requires a set of representative samples—the so-called model set. The model set should contain samples that span the sources of expected variability during the model's maintenance [48].

In certain situations it is possible to design a model set and prepare samples according to the principles of experimental design (DoE). However, such an approach cannot be used in drug authenticity studies. Alternatively, the model set is selected from the available set of samples that represent authentic and counterfeit drugs. This can be done, for instance, by drawing a number of samples randomly using clustering methods (e.g., k-means, SOM), optimality criteria (D-optimality) and by assuming a uniform design of the model set samples. The latter approach seems to be the most popular. The Kennard and Stone algorithm can be used to ensure a nearly uniform distribution of samples [48]. In classification/discrimination problems, the selection of the representative model set should be carried out individually for each group of samples. Depending on the method considered, a balanced representation of samples (containing the same number of samples from each group) may be required to ensure the optimal construction of a logic rule (for instance PLS-DA).

3.3.4. Selection of relevant variables

A large number of redundant variables, in general, increases the risk of model overfitting, affects the predictive performance and makes their interpretation more complex. This is why variable selection is often considered to be an additional step in data modeling. Many variable selection methods have been proposed in the literature. Some of them were developed specifically for a given modeling approach. For instance, variable importance in projection (VIP) [49] and significance multivariate correlation (SMC) [50] were proposed for PLS. Variable selection based on optimization frameworks using, for instance, generic algorithms, is not restricted to a particular classification or discriminant method. A comprehensive overview of variable selection techniques can be found in reference [51].

3.3.5. Validation of discriminant/classification models

Validation of a model is a crucial step. In general, there is a distinction between internal and external model validation whose aim is to evaluate performance of a model with respect to its predictive abilities [52]. Usually, internal validation methods (re-sampling approaches) are applied to verify the complexity of a model and thus to minimize the risk of its overfitting. Internal validation methods follow certain data sampling strategies, for instance, leave-one-out cross validation, leave-M-out cross validation, V-fold cross validation, Monte Carlo cross validation, jackknifing and bootstrapping. External validation examines its predictive performance using an independent set of test set samples. It is important to stress that these samples do not enter any stage of the model's construction. That is why they are regarded as an independent test set. Subset selection approaches, which were mentioned in the Section 3.3.3, split available samples into a model set and an external test set. When the uniform representation of test set samples is important, a modified variant of the Kennard and Stone algorithm, the so-called 'Duplex' algorithm [48], should be considered.

Various figures of merit are used to score discrimination or classification models, for instance, the correct classification rate (CCR), sensitivity (true positive rate), specificity (true negative rate), efficiency, accuracy, precision or area under curve (AUC).

These measures can be calculated in the course of any resampling method for the model and test set samples, e.g. the Monte Carlo validation [53].

4. Selected examples of case studies and discussion

Based on a literature survey and the selected examples of case studies listed in Table 1, it can be concluded that chromatographic-based techniques and spectroscopic methods such as Raman, FT-IR, mid-IR and NIR spectroscopy are frequently selected to evaluate drug samples. Analysis of published case studies provides readers with substantial evidence that leads to the conclusion that the detection of counterfeit medicines, in most of cases, is a multivariate problem. Often, drug samples, regardless of their type and form, are characterized by chromatographic or spectroscopic fingerprints that can lead to high-dimensional data. Therefore, the use of chemometric techniques to explore such complex data and to obtain adequate diagnostic models is inevitable. Depending on the problem at hand, the selection of an analytical approach and the appropriate preprocessing followed by the choice of suitable chemometric method(s) should be optimized. In the context of data preprocessing and further chemometric data analysis, it is difficult to formulate general guidelines regarding the selection of the appropriate techniques. It is easier to exclude certain methods a priori based on their objectives than to conclude beforehand that a certain discriminant or classification method that belongs to the same family will perform considerably better—except in some obvious situations. For instance, linear techniques are suitable for handling linear problems, whereas non-linear techniques should be used only for non-linear problems. A rational strategy for data preprocessing and modeling favors the use of the simplest available and the least complex method (with the smallest possible number of latent factors, variables, etc.) in order to obtain a model with acceptable figures of merit (e.g., sensitivity, specificity, correct classification rate, etc.) and that is in agreement with the general knowledge. Therefore, linear data modeling methods are used initially. When they fail in terms of satisfactory predictive abilities, one usually tries to determine the potential causes. A general troubleshooting procedure focuses on the improvement of a diagnostic model through the enhancement of the signal-to-noise ratio, the suppression of undesired and unrelated effect(s) (e.g. peak shifts in chromatograms, elimination of scattering), the identification and elimination of potential outlying samples or the elimination of redundant variables.

Not surprisingly, PCA and HCA methods are used to visualize the structure of the multivariate data in the majority of studies. To distinguish between authentic and counterfeit drug samples diagnostic models including discriminant and classification models are built, e.g. PLS-DA k-NN and SIMCA. The choice between a discriminant and classification diagnostic model should be made carefully, especially when verification of a drug's authenticity is the major concern [54]. Classification methods are the only choice when the variability of counterfeit samples cannot be sufficiently sampled. In that case, the construction of logic rule(s) using the two-class discriminant model may lead to the incorrect recognition of new samples if all of the relevant sources of variability are not incorporated during model's construction. On the other hand, when the potential variability of counterfeit samples has a limit, for instance, when the amount of the impurities found in counterfeit samples is always larger than in the genuine samples, discriminant techniques can also be considered as an option [53].

5. Conclusions

In this review article, we discuss a number of applications of the most popular analytical and chemometric methods that are used to identify counterfeit drugs. Detection of counterfeit medicines is certainly an extremely challenging problem from the analytical and chemometric point of view. The specific chemical targets that can serve as potential markers of the counterfeiting process vary from case to case. This is why an untargeted approach that is based on developing and then modeling chromatographic or spectroscopic fingerprints seems to be the most frequently selected strategy. Apparently, the combined use of appropriately selected instrumental techniques, which are able to collect the relevant analytical data to describe a sample, and chemometric methods support the task of drug authentication to a great extent.

In our opinion, one of the major challenges in such studies is related to the sampling of the expected sources of variability. In general, the group that is represented by authentic samples is relatively compact because variation is strictly controlled in the course of the production process. Therefore, sampling authentic drugs is a rather easy task because different batches of samples are available directly from a producer. Unfortunately, the variability in the group of counterfeit medicines is much larger and hence the sampling of all possible sources of variability is hardly possible. This is why, one-class classifiers that focus on modeling only one group of samples are considered most frequently.

Another important issue is related to the proper validation of the diagnostic models. Unfortunately, this aspect is underestimated in many studies. Due to the limited number of samples, the constructed models are validated using different variants of cross validation, although these validation approaches, in general, cannot substitute for validation that is based on an external test set.

References

[1] K. Dégardin, Y. Roggo, P. Margot, Understanding and fighting the medicine counterfeit market, J. Pharm. Biomed. Anal. 87 (2014) 167–175, http://dx.doi.org/10.1016/j.jpba.2013.01.009.
[2] P.-Y. Sacré, E. Deconinck, T. De Beer, P. Courselle, R. Vancauwenberghe, P. Chiap, et al., Comparison and combination of spectroscopic techniques for the detection of counterfeit medicines, J. Pharm. Biomed. Anal. 53 (2010) 445–453, http://dx.doi.org/10.1016/j.jpba.2010.05.012.
[3] O.Y. Rodionova, L.P. Houmøller, A.L. Pomerantsev, P. Geladi, J. Burger, V.L. Dorofeyev, et al., NIR spectrometry for counterfeit drug detection: a feasibility study, Anal. Chim. Acta. 549 (2005) 151–158, http://dx.doi.org/10.1016/j.aca.2005.06.018.
[4] R. Martino, M. Malet-Martino, V. Gilard, S. Balayssac, Counterfeit drugs: analytical techniques for their identification, Anal. Bioanal Chem. 398 (2010) 77–92, http://dx.doi.org/10.1007/s00216-010-3748-y.
[5] M.T. Koesdjojo, Y. Wu, A. Boonloed, E.M. Dunfield, V.T. Remcho, Low-cost, high-speed identification of counterfeit antimalarial drugs on paper, Talanta 130 (2014) 122–127, http://dx.doi.org/10.1016/j.talanta.2014.05.050.
[6] S. Wilczyński, The use of dynamic thermal analysis to distinguish between genuine and counterfeit drugs, Int. J. Pharm. 490 (2015) 16–21, http://dx.doi.org/10.1016/j.ijpharm.2015.04.077.
[7] N. Lindegårdh, T.T. Hien, J. Farrar, P. Singhasivanon, N.J. White, N.P.J. Day, A simple and rapid liquid chromatographic assay for evaluation of potentially counterfeit Tamiflu®, J. Pharm. Biomed. Anal. 42 (2006) 430–433, http://dx.doi.org/10.1016/j.jpba.2006.04.028.
[8] A. Panusa, G. Multari, G. Incarnato, L. Gagliardi, High-performance liquid chromatography analysis of anti-inflammatory pharmaceuticals with ultraviolet and electrospray-mass spectrometry detection in suspected counterfeit homeopathic medicinal products, J. Pharm. Biomed. Anal. 43 (2007) 1221–1227, http://dx.doi.org/10.1016/j.jpba.2006.10.012.
[9] E. Deconinck, M. Canfyn, P.-Y. Sacré, S. Baudewyns, P. Courselle, J.O. De Beer, A validated GC–MS method for the determination and quantification of residual solvents in counterfeit tablets and capsules, J. Pharm. Biomed. Anal. 70 (2012) 64–70, http://dx.doi.org/10.1016/j.jpba.2012.05.022.
[10] R.D. Marini, E. Rozet, M.L.A. Montes, C. Rohrbasser, S. Roht, D. Rhème, et al., Reliable low-cost capillary electrophoresis device for drug quality control and counterfeit medicines, J. Pharm. Biomed. Anal. 53 (2010) 1278–1287, http://dx.doi.org/10.1016/j.jpba.2010.07.026.
[11] C. Ricci, L. Nyadong, F. Yang, F.M. Fernandez, C.D. Brown, P.N. Newton, et al., Assessment of hand-held Raman instrumentation for in situ screening for potentially counterfeit artesunate antimalarial tablets by FT-Raman

spectroscopy and direct ionization mass spectrometry, Anal. Chim. Acta 623 (2008) 178–186, http://dx.doi.org/10.1016/j.aca.2008.06.007.
[12] R. Santamaria-Fernandez, R. Hearn, J.-C. Wolff, Detection of counterfeit antiviral drug Heptodin™ and classification of counterfeits using isotope amount ratio measurements by multicollector inductively coupled plasma mass spectrometry (MC-ICPMS) and isotope ratio mass spectrometry (IRMS), Sci. Justice 49 (2009) 102–106, http://dx.doi.org/10.1016/j.scijus.2008.12.003.
[13] V. Silvestre, V.M. Mboula, C. Jouitteau, S. Akoka, R.J. Robins, G.S. Remaud, Isotopic 13C NMR spectrometry to assess counterfeiting of active pharmaceutical ingredients: site-specific 13C content of aspirin and paracetamol, J. Pharm. Biomed. Anal. 50 (2009) 336–341, http://dx.doi.org/10.1016/j.jpba.2009.04.030.
[14] L. Komsta, M. Waksmundzka-Hajnos, J. Sherma, Thin Layer Chromatography in Drug Analysis, CRC Press, 2013, 2016.
[15] M.E. ElTantawy, L.I. Bebawy, R.F. Shokry, Chromatographic determination of clopidogrel bisulfate; detection and quantification of counterfeit Plavix® tablets, Bull. Fac. Pharm. Cairo Univ. 52 (2014) 91–101, http://dx.doi.org/10.1016/j.bfopcu.2014.04.003.
[16] D.B. da Justa Neves, R.G.A. Marcheti, E.D. Caldas, Incidence of anabolic steroid counterfeiting in Brazil, Forensic Sci. Int. 228 (2013) e81–e83, http://dx.doi.org/10.1016/j.forsciint.2013.02.035.
[17] J. Luypaert, D.L. Massart, Y. Vander Heyden, Near-infrared spectroscopy applications in pharmaceutical analysis, Talanta 72 (2007) 865–883, http://dx.doi.org/10.1016/j.talanta.2006.12.023.
[18] S. Wartewig, R. Neubert, Pharmaceutical applications of Mid-IR and Raman spectroscopy, Adv. Drug Deliv Rev. 57 (2005) 1144–1170, http://dx.doi.org/10.1016/j.addr.2005.01.022.
[19] U. Holzgrabe, M. Malet-Martino, Analytical challenges in drug counterfeiting and falsification—the NMR approach, J. Pharm. Biomed. Anal. 55 (2011) 679–687, http://dx.doi.org/10.1016/j.jpba.2010.12.017.
[20] K. Dégardin, Y. Roggo, F. Been, P. Margot, Detection and chemical profiling of medicine counterfeits by Raman spectroscopy and chemometrics, Anal. Chim. Acta 705 (2011) 334–341, http://dx.doi.org/10.1016/j.aca.2011.07.043.
[21] I. McEwen, A. Elmsjö, A. Lehnström, B. Hakkarainen, M. Johansson, Screening of counterfeit corticosteroid in creams and ointments by NMR spectroscopy, J. Pharm. Biomed. Anal. 70 (2012) 245–250, http://dx.doi.org/10.1016/j.jpba.2012.07.005.
[22] U. Holzgrabe, R. Deubner, C. Schollmayer, B. Waibel, Quantitative NMR spectroscopy—applications in drug analysis, J. Pharm. Biomed. Anal. 38 (2005) 806–812, http://dx.doi.org/10.1016/j.jpba.2005.01.050.
[23] J. Trygg, J. Gabrielsson, T. Lundstedt, Background estimation, denoising, and preprocessing, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Oxford, 2009, pp. 1–8 (accessed 21.01.15) http://www.sciencedirect.com/science/article/pii/B9780444527011000971.
[24] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, et al., Breaking with trends in pre-processing? TrAC Trends Anal. Chem. 50 (2013) 96–106, http://dx.doi.org/10.1016/j.trac.2013.04.015.
[25] P.H.C. Eilers, Baseline correction with asymmetric least squares smoothing, Anal. Chem. 75 (2003) 3631–3636.
[26] A.M. van Nederkassel, M. Daszykowski, P.H.C. Eilers, Y. Vander Heyden, A comparison of three algorithms for chromatograms alignment, J. Chromatogr. A. 1118 (2006) 199–210.
[27] M. Daszykowski, Y. Vander Heyden, C. Boucon, B. Walczak, Automated alignment of one-dimensional chromatographic fingerprints, J. Chromatogr. A. 1217 (2010) 6127–6133.
[28] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemom. 16 (2002) 119–128, http://dx.doi.org/10.1002/cem.695.
[29] P. Geladi, D. MacDougall, H. Martens, Linearization and scatter-correction for near-infrared reflectance spectra of meat, Appl. Spectrosc. 39 (1985) 491–500, http://dx.doi.org/10.1366/0003702854248656.
[30] O.M. Kvalheim, F. Brakstad, Y. Liang, Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise, Anal. Chem. 66 (1994) 43–51, http://dx.doi.org/10.1021/ac00073a010.
[31] B. Massart, Q. Guo, F. Questier, D.L. Massart, C. Boucon, S. de Jong, et al., Data structures and data transformations for clustering chemical data, TrAC Trends Anal. Chem. 20 (2001) 35–41, http://dx.doi.org/10.1016/S0167-2940(01)90097-4.
[32] I. Stanimirova, B. Walczak, D.L. Massart, Multiple factor analysis in environmental chemistry, Anal. Chim. Acta 545 (2005) 1–12, http://dx.doi.org/10.1016/j.aca.2005.04.054.
[33] M. Daszykowski, B. Walczak, D.L. Massart, Projection methods in chemistry, Chemom. Intell. Lab. Syst. 65 (2003) 97–112, http://dx.doi.org/10.1016/S0169-7439(02)00107-7.
[34] D.L. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Robert E. Krieger Publishing Company, Malabar, Florida, 1989.
[35] R. Todeschini, D. Ballabio, V. Consonni, Distances and other dissimilarity measures in chemometrics, in: Encyclopedia of Analytical Chemistry, John Wiley & Sons, Ltd, 2006 (accessed 13.04.15) http://onlinelibrary.wiley.com/doi/10.1002/9780470027318.a9438/abstract.
[36] M. Daszykowski, B. Walczak, D.L. Massart, On the optimal partitioning of data with K-means, growing K-means, neural gas, and growing neural gas, J. Chem. Info. Comput. Sci. 42 (2002) 1378–1389.
[37] R.G. Brereton, Pattern recognition in chemometrics, Chemom. Intell. Lab. Syst. (2015), http://dx.doi.org/10.1016/j.chemolab.2015.06.012.
[38] M. Forina, P. Oliveri, S. Lanteri, M. Casale, Class-modeling techniques, classic and new, for old and new problems, Chemom. Intell. Lab. Syst. 93 (2008) 132–148, http://dx.doi.org/10.1016/j.chemolab.2008.05.003.
[39] M. Forina, C. Armanino, R. Leardi, G. Drava, A class-modelling technique based on potential functions, J. Chemom. 5 (1991) 435–453, http://dx.doi.org/10.1002/cem.1180050504.
[40] M.P. Derde, D.L. Massart, UNEQ: a class modelling supervised pattern recognition technique, Mikrochim. Acta 89 (1986) 139–152, http://dx.doi.org/10.1007/BF01207313.
[41] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth International Group, Belmont, CA, 1984.
[42] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away, J. Chemom. 28 (2014) 213–225, http://dx.doi.org/10.1002/cem.2609.
[43] D. Svozil, V. Kvasnicka, J. Pospichal, Introduction to multi-layer feed-forward neural networks, Chemom. Intell. Lab. Syst. 39 (1997) 43–62, http://dx.doi.org/10.1016/S0169-7439(97)00061-0.
[44] A.I. Belousov, S.A. Verzakov, J. von Frese, Applicational aspects of support vector machines, J. Chemom. 16 (2002) 482–489, http://dx.doi.org/10.1002/cem.744.
[45] R.G. Brereton, One-class classifiers, J. Chemom. 25 (2011) 225–246, http://dx.doi.org/10.1002/cem.1397.
[46] L. Xu, M. Goodarzi, W. Shi, C.-B. Cai, J.-H. Jiang, A MATLAB toolbox for class modeling using one-class partial least squares (OCPLS) classifiers, Chemom. Intell. Lab. Syst. 139 (2014) 58–63, http://dx.doi.org/10.1016/j.chemolab.2014.09.005.
[47] R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Pavan, CAIMAN (Classification And Influence Matrix Analysis): a new approach to the classification based on leverage-scaled functions, Chemom. Intell. Lab. Syst. 87 (2007) 3–17, http://dx.doi.org/10.1016/j.chemolab.2005.11.001.
[48] M. Daszykowski, B. Walczak, D.L. Massart, Representative subset selection, Anal. Chim. Acta 468 (2002) 91–103, http://dx.doi.org/10.1016/S0003-2670(02)00651-7.
[49] O.M. Kvalheim, R. Arneberg, O. Bleie, T. Rajalahti, A.K. Smilde, J.A. Westerhuis, Variable importance in latent variable regression models, J. Chemom. 28 (2014) 615–622, http://dx.doi.org/10.1002/cem.2626.
[50] T.N. Tran, N.L. Afanador, L.M.C. Buydens, L. Blanchet, Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC), Chemom. Intell. Lab. Syst. 138 (2014) 153–160, http://dx.doi.org/10.1016/j.chemolab.2014.08.005.
[51] I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh, Feature Extraction, Foundations and Applications, Springer, Berlin, 2006.
[52] K.H. Esbensen, P. Geladi, Principles of proper validation: use and abuse of re-sampling for validation, J. Chemom. 24 (2010) 168–187, http://dx.doi.org/10.1002/cem.1310.
[53] B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles, Analyst 141 (2016) 1060–1070, http://dx.doi.org/10.1039/C5AN01656H.
[54] O.Y. Rodionova, A.V. Titova, A.L. Pomerantsev, Discriminant analysis is an inappropriate method of authentication, TrAC Trends Anal. Chem. 78 (2016) 17–22, http://dx.doi.org/10.1016/j.trac.2016.01.010.
[55] F.E. Dowell, E.B. Maghirang, F.M. Fernandez, P.N. Newton, M.D. Green, Detecting counterfeit antimalarial tablets by near-infrared spectroscopy, J. Pharm. Biomed. Anal. 48 (2008) 1011–1014, http://dx.doi.org/10.1016/j.jpba.2008.06.024.
[56] K. Kwok, L.S. Taylor, Analysis of counterfeit Cialis® tablets using Raman microscopy and multivariate curve resolution, J. Pharm. Biomed. Anal. 66 (2012) 126–135, http://dx.doi.org/10.1016/j.jpba.2012.03.026.
[57] D. Custers, B. Krakowska, J.O. De Beer, P. Courselle, M. Daszykowski, S. Apers, et al., Chromatographic impurity fingerprinting of genuine and counterfeit Cialis® as a means to compare the discriminating ability of PDA and MS detection, Talanta 146 (2016) 540–548, http://dx.doi.org/10.1016/j.talanta.2015.09.029.
[58] T. Puchert, D. Lochmann, J.C. Menezes, G. Reich, Near-infrared chemical imaging (NIR-CI) for counterfeit drug identification—a four-stage concept with a novel approach of data processing (Linear Image Signature), J. Pharm. Biomed. Anal. 51 (2010) 138–145, http://dx.doi.org/10.1016/j.jpba.2009.08.022.
[59] R. da Silva Fernandes, F.S.L. da Costa, P. Valderrama, P.H. Março, K.M.G. de Lima, Non-destructive detection of adulterated tablets of glibenclamide using NIR and solid-phase fluorescence spectroscopy and chemometric methods, J. Pharm. Biomed. Anal. 66 (2012) 85–90, http://dx.doi.org/10.1016/j.jpba.2012.03.004.
[60] M.B. Lopes, J.-C. Wolff, Investigation into classification/sourcing of suspect counterfeit Heptodin™ tablets by near infrared chemical imaging, Anal. Chim. Acta 633 (2009) 149–155, http://dx.doi.org/10.1016/j.aca.2008.11.036.
[61] M.B. Lopes, J.-C. Wolff, J.M. Bioucas-Dias, M.A.T. Figueiredo, Determination of the composition of counterfeit Heptodin tablets by near infrared chemical imaging and classical least squares estimation, Anal. Chim. Acta 641 (2009) 46–51, http://dx.doi.org/10.1016/j.aca.2009.03.034.
[62] F. Lu, X. Weng, Y. Chai, Y. Yang, Y. Yu, G. Duan, A novel identification system for counterfeit drugs based on portable Raman spectroscopy, Chemom. Intell. Lab. Syst. 127 (2013) 63–69, http://dx.doi.org/10.1016/j.chemolab.2013.06.001.

[63] P. de Peinder, M.J. Vredenbregt, T. Visser, D. de Kaste, Detection of Lipitor® counterfeits: a comparison of NIR and Raman spectroscopy in combination with chemometrics, J. Pharm. Biomed. Anal. 47 (2008) 688–694, http://dx.doi.org/10.1016/j.jpba.2008.02.016.
[64] M. Dumarey, A.M. van Nederkassel, I. Stanimirova, M. Daszykowski, F. Bensaid, M. Lees, et al., Recognizing paracetamol formulations with the same synthesis pathway based on their trace-enriched chromatographic impurity profiles, Anal. Chim. Acta 655 (2009) 43–51, http://dx.doi.org/10.1016/j.aca.2009.09.050.
[65] O.Y. Rodionova, A.L. Pomerantsev, NIR-based approach to counterfeit-drug detection, TrAC Trends Anal. Chem. 29 (2010) 795–803, http://dx.doi.org/10.1016/j.trac.2010.05.004.
[66] M. de Veij, A. Deneckere, P. Vandenabeele, D. de Kaste, L. Moens, Detection of counterfeit Viagra® with Raman spectroscopy, J. Pharm. Biomed. Anal. 46 (2008) 303–309, http://dx.doi.org/10.1016/j.jpba.2007.10.021.
[67] P.-Y. Sacré, E. Deconinck, L. Saerens, T. De Beer, P. Courselle, R. Vancauwenberghe, et al., Detection of counterfeit Viagra® by Raman microspectroscopy imaging and multivariate analysis, J. Pharm. Biomed. Anal. 56 (2011) 454–461, http://dx.doi.org/10.1016/j.jpba.2011.05.042.
[68] D. Custers, T. Cauwenbergh, J.L. Bothy, P. Courselle, J.O. De Beer, S. Apers, et al., ATR-FTIR spectroscopy and chemometrics: an interesting tool to discriminate and characterize counterfeit medicines, J. Pharm. Biomed. Anal. 112 (2015) 181–189, http://dx.doi.org/10.1016/j.jpba.2014.11.007.
[69] M.J. Anzanello, R.S. Ortiz, R.P. Limbergerb, P. Mayorga, A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes, J. Pharm. Biomed. Anal. 83 (2013) 209–214, http://dx.doi.org/10.1016/j.jpba.2013.05.004.
[70] P.-Y. Sacré, E. Deconinck, M. Daszykowski, P. Courselle, R. Vancauwenberghe, P. Chiap, et al., Impurity fingerprints for the identification of counterfeit medicines—a feasibility study, Anal. Chim. Acta 701 (2011) 224–231, http://dx.doi.org/10.1016/j.aca.2011.05.041.
[71] E. Deconinck, P.Y. Sacré, P. Courselle, J.O. De Beer, Chemometrics and chromatographic fingerprints to discriminate and classify counterfeit medicines containing PDE-5 inhibitors, Talanta 100 (2012) 123–133, http://dx.doi.org/10.1016/j.talanta.2012.08.029.
[72] E. Deconinck, P.Y. Sacré, D. Coomans, J. De Beer, Classification trees based on infrared spectroscopic data to discriminate between genuine and counterfeit medicines, J. Pharm. Biomed. Anal. 57 (2012) 68–75, http://dx.doi.org/10.1016/j.jpba.2011.08.036.
[73] C.R. Jung, R.S. Ortiz, R. Limberger, P. Mayorga, A new methodology for detection of counterfeit Viagra® and Cialis® tablets by image processing and statistical analysis, Forensic Sci. Int. 216 (2012) 92–96, http://dx.doi.org/10.1016/j.forsciint.2011.09.002.
[74] R.S. Ortiz, K.C. de Mariotti, M.H. Holzschuh, W. Romão, R.P. Limberger, P. Mayorga, Profiling counterfeit Cialis, Viagra and analogs by UPLC–MS, Forensic Sci. Int. 229 (2013) 13–20, http://dx.doi.org/10.1016/j.forsciint.2013.03.024.
[75] R.S. Ortiz, K.C. Mariotti, N.V. Schwab, G.P. Sabin, W.F.C. Rocha, E.V.R. de Castro, et al., Fingerprinting of sildenafil citrate and tadalafil tablets in pharmaceutical formulations via X-ray fluorescence (XRF) spectrometry, J. Pharm. Biomed. Anal. 58 (2012) 7–11, http://dx.doi.org/10.1016/j.jpba.2011.09.005.

Brussels, 21/06/2016

Deborah Custers, PhD
Scientific Institute of Public Division of Food
Medicines and Consumer Safety
Section Medicinal Products Health (WIV-ISP)
Juliette Wytsmanstraat 14
Brussels
Belgium

To whom it may concern

Hereby,

I declare that my overall contribution to the article entitled "Chemometrics and identification of counterfeit medicines – a review" by B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, published in Journal of Pharmaceutical and Biomedical Analysis, 127 (2016) 112-122 my contribution mostly concerned:

- assistance in explaining the medicines' counterfeiting issue from a pharmaceutical point of view,
- discussion regarding the content of the review paper – selection of analytical methods,
- providing comments on the manuscript.

Deborah Custers

21-06-2016

WETENSCHAPPELIJK INSTITUUT
VOLKSGEZONDHEID
INSTITUT SCIENTIFIQUE
DE SANTÉ PUBLIQUE

Section of Medicines

datum : 21/06/2016
uw ref. :
onze ref.

contact : Eric Deconinck
tel. : + 32 2 642 51 70
fax : + 32 2 642 53 27
e-mail : Eric Deconinck@wiv-isp.be

To whom it may concern

Hereby,

I declare that my overall contribution to the article entitled "Chemometrics and identification of counterfeit medicines – a review" by B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, published in Journal of Pharmaceutical and Biomedical Analysis, 127 (2016) 112-122 mostly concerned:

- assistance in explaining counterfeiting issue from a pharmaceutical point of view,
- providing comments on the manuscript,
- participating in preparing answers to reviewers' and assistance in preparing a revised version of the manuscript.

Pharm. Eric Deconinck, PhD
Head of section
Section of medicines
Operational Direction Food
Medicines and Consumer Safety

Juliette Wytsmanstraat 14
1050 Brussel | België
T + 32 2 642 51 11 | F + 32 2 642 50 01
info@iph.fgov.be | www.iph.fgov.be

dr hab. Michał Daszykowski, prof. UŚ
Instytut Chemii
Uniwersytet Śląski
ul. Szkolna 9
40-006 Katowice

Katowice 29.06.2016

Oświadczam, że w artykule pt. „Chemometrics and identification of counterfeit medicines - a review" opublikowanym w czasopiśmie Journal of Pharmaceutical and Biomedical Analysis 127 (2016) 112-122 mój udział polegał na:

- współtworzeniu ogólnej koncepcji artykułu,
- pomocy w doborze pozycji literaturowych,
- weryfikacji wniosków uzyskanych przez doktorantkę,
- opiece i merytorycznym nadzorze procesu przygotowania manuskryptu,
- pomocy w przeprowadzeniu procedury redakcyjnej i przygotowaniu odpowiedzi na recenzje,
- dokonaniu ostatecznej korekty artykułu.

Michał Daszykowski